2014

A reference genome for common bean and genome-wide analysis of dual domestications

Jeremy Schmutz
US Department of Energy Joint Genome Institute, Walnut Creek, California

Phillip E. McClean
North Dakota State University

Sujan Mamidi
North Dakota State University

G. Albert Wu
US Department of Energy Joint Genome Institute, Walnut Creek, California

Steven B. Cannon
4Corn Insects and Crop Genetics Research Unit, US Department of Agriculture–Agricultural Research Service, Ames, Iowa

See next page for additional authors

Authors
Jeremy Schmutz, Phillip E. McClean, Sujan Mamidi, G. Albert Wu, Steven B. Cannon, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, Qijian Song, Carolina Chavarro, Mirayda Torres-Torres, Valerie Geffroy, Samira Mafi Moghaddam, Dongying Gao, Brian Abernathy, Kerrie Barry, Matthew Blair, Mark A. Brick, Mansi Chovatia, Paul Gepts, David M. Goodstein, Michael Gonzales, Uffe Hellsten, D. L. Hyten, Gaofeng Jia, James D. Kelly, Dave Kudrna, Rian Lee, Manon M.S. Richard, Phillip N. Miklas, Juan M. Osorno, Josiane Rodrigues, Vincent Thareau, Carlos A. Urrea Florez, Mei Wang, Yeisoo Yu, Ming Zhang, Rod A. Wing, P. B. Cregan, Daniel S. Rokhsar, and Scott A. Jackson

nature genetics
OPEN

A reference genome for common bean and genome-wide analysis of dual domestications

Jeremy Schmutz[1,2,17], Phillip E McClean[3,17], Sujan Mamidi[3], G Albert Wu[1], Steven B Cannon[4], Jane Grimwood[2], Jerry Jenkins[2], Shengqiang Shu[1], Qijian Song[5], Carolina Chavarro[6], Mirayda Torres-Torres[6], Valerie Geffroy[7,8], Samira Mafi Moghaddam[3], Dongying Gao[6], Brian Abernathy[6], Kerrie Barry[1], Matthew Blair[9], Mark A Brick[10], Mansi Chovatia[1], Paul Gepts[11], David M Goodstein[1], Michael Gonzales[6], Uffe Hellsten[1], David L Hyten[5,16], Gaofeng Jia[5], James D Kelly[12], Dave Kudrna[13], Rian Lee[3], Manon M S Richard[7], Phillip N Miklas[14], Juan M Osorno[3], Josiane Rodrigues[5,16], Vincent Thareau[7], Carlos A Urrea[15], Mei Wang[1], Yeisoo Yu[13], Ming Zhang[1], Rod A Wing[13], Perry B Cregan[5], Daniel S Rokhsar[1] & Scott A Jackson[6]

Common bean (*Phaseolus vulgaris* L.) is the most important grain legume for human consumption and has a role in sustainable agriculture owing to its ability to fix atmospheric nitrogen. We assembled 473 Mb of the 587-Mb genome and genetically anchored 98% of this sequence in 11 chromosome-scale pseudomolecules. We compared the genome for the common bean against the soybean genome to find changes in soybean resulting from polyploidy. Using resequencing of 60 wild individuals and 100 landraces from the genetically differentiated Mesoamerican and Andean gene pools, we confirmed 2 independent domestications from genetic pools that diverged before human colonization. Less than 10% of the 74 Mb of sequence putatively involved in domestication was shared by the two domestication events. We identified a set of genes linked with increased leaf and seed size and combined these results with quantitative trait locus data from Mesoamerican cultivars. Genes affected by domestication may be useful for genomics-enabled crop improvement.

Common bean (*P. vulgaris* L.) is a crop of major societal importance and is a major source of protein and essential nutrients. Worldwide, common bean is the most consumed legume, providing up to 15% of total daily calories and 36% of total daily protein in parts of Africa and the Americas (see URLs). More than 200 million people in sub-Saharan Africa depend on the common bean as a primary staple. It has many health-beneficial[1,2] nutrients whose concentrations are heritable[3], and increasing the concentrations of these nutrients is a breeding objective worldwide[4].

Multiple lines of evidence have shown that wild common bean is organized in two geographically isolated and genetically differentiated wild gene pools (Mesoamerican and Andean) that diverged from a common ancestral wild population more than 100,000 years ago[5]. From these wild gene pools, nearly 8,000 years ago, common bean was independently domesticated in what is now Mexico and in South America[6–9], and these domestication events were followed by local adaptations resulting in landraces with distinct characteristics. In what is now Mexico, common bean was likely domesticated concurrently with maize as part of the 'milpa' cropping system (featuring common bean along with maize and squash), which was adopted throughout the Americas[10]. Domestication led to morphological changes, including increased seed and leaf sizes, changes in growth habit and photoperiod responses[11], and variation in seed coat color and pattern that distinguish culturally adapted classes of beans[12].

Independent domestication events, starting from distinct gene pools of a single species, provide experimental replication not typically found in domestication or evolutionary studies. It is possible to deduce domestication history on a genome-wide scale and examine the roles of parallel evolution and introgression during the domestication of two independent lineages within a single species. Here, to understand

[1]US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [2]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [3]Department of Plant Sciences, North Dakota State University, Fargo, North Dakota, USA. [4]Corn Insects and Crop Genetics Research Unit, US Department of Agriculture–Agricultural Research Service, Ames, Iowa, USA. [5]Soybean Genomics and Improvement Laboratory, US Department of Agriculture–Agricultural Research Service, Beltsville, Maryland, USA. [6]Center for Applied Genetic Technologies, University of Georgia, Athens, Georgia, USA. [7]CNRS, Université Paris–Sud, Institut de Biologie des Plantes, UMR 8618, Saclay Plant Sciences (SPS), Orsay, France. [8]Institut National de la Recherche Agronomique (INRA), Université Paris–Sud, Unité Mixte de Recherche de Génétique Végétale, Gif-sur-Yvette, France. [9]Department of Agricultural and Natural Sciences, Tennessee State University, Nashville, Tennessee, USA. [10]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado, USA. [11]Department of Plant Sciences, University of California, Davis, Davis, California, USA. [12]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan, USA. [13]Arizona Genomics Institute, University of Arizona, Tucson, Arizona, USA. [14]Vegetable and Forage Crop Research Unit, US Department of Agriculture–Agricultural Research Service, Prosser, Washington, USA. [15]Panhandle Research and Extension Center, University of Nebraska, Scottsbluff, Nebraska, USA. [16]Present addresses: Pioneer Hi-Bred International, Inc., Johnston, Iowa, USA (D.L.H.) and Genética e Melhoramento, Federal University of Viçosa, Viçosa, Brazil (J.R.). [17]These authors contributed equally to this work. Correspondence should be addressed to S.A.J. (sjackson@uga.edu), J.S. (jschmutz@hudsonalpha.org) or P.E.M. (phillip.mcclean@ndsu.edu).

Received 8 November 2013; accepted 15 May 2014; published online 8 June 2014; doi:10.1038/ng.3008

the history of these complicated domestication events and their implications for modern bean crop improvement, we report a genome sequence for an Andean ecotype of common bean and an analysis of genetic variation in accessions ranging from Mexico to the southern range of the species in Argentina. In addition, comparative genomics with soybean (*Glycine max*), a closely related crop, identified effects of shared and lineage-dependent polyploidies on gene fractionation and recent transposable element expansion in the common bean.

RESULTS

Reference genome and analysis

To obtain a high-quality reference genome, we sequenced an inbred landrace line of *P. vulgaris* (G19833) derived from the Andean pool (Race Peru) using a whole-genome shotgun sequencing strategy that combined multiple linear libraries (18.6× assembled sequence coverage) and ten paired libraries of varying insert sizes (1.8× assembled) sequenced with the Roche 454 platform together with 24.1 Gb of Illumina-sequenced fragment libraries. For longer-range linkage, we also end sequenced three fosmid libraries and two BAC libraries on the Sanger platform (0.54× long-insert pairs) for a total assembled sequence coverage level of 21.0× (**Supplementary Tables 1** and **2**). The resulting assembled sequences were organized into 11 chromosomal pseudomolecules by integration with a dense GoldenGate- and Infinium-based SNP map of 7,015 markers typed on 267 F_2 lines from a Stampede × Red Hawk cross and a similar set of Infinium markers and 261 SSRs (simple sequence repeats) typed on 88 F_5-derived recombinant inbred lines (RILs) derived from the same cross (P.B.C. and Q.S., unpublished data). Additional refinements to the pseudomolecules were made on the basis of synteny with soybean (*G. max*), where allowed by available map data. Almost all of these changes were made in pericentromeric regions, where recombination is generally too limited to resolve the ordering and orientation of small scaffolds. The pseudomolecules included 468.2 Mb of mapped sequence in 240 scaffolds. The total release includes 472.5 Mb of the ~587-Mb genome (see URLs), with half of the assembled nucleotides in contigs longer than 39.5 kb (contig N50) (**Supplementary Table 3**). To annotate the chromosomal assembly, we combined Sanger-derived EST resources and a substantial amount of new RNA sequencing (RNA-seq) reads (727 million reads from 11 tissues and developmental stages; **Supplementary Table 4**) with homology-based and *de novo* gene prediction approaches. The resulting annotation includes 27,197 protein-coding loci, including 4,491 alternative transcripts (**Supplementary Table 5**), an underestimate that will increase with additional transcriptomes and analyses. Most of these genes (91%) were retained in synteny blocks with *G. max* (**Supplementary Note**).

We identified recent transposable element activity and expansions of transposon numbers (**Supplementary Figs. 1–3**). Although recently diverged repeats could not be annotated directly from Roche 454 pyrosequencing data, extensive BAC-end and fosmid-end sequence data and a dense genetic map allowed us to position 99.6% of genic sequences and to link into those genes embedded in regions dense with transposable elements (**Supplementary Figs. 4–14**). Centromere and pericentromeric regions were primarily repetitive, and, similar to in other sequenced genomes[13,14], these pericentromeric genomic regions were recombinationally inert (**Supplementary Fig. 15** and **Supplementary Table 6**). Using a threshold of 2 Mb/cM to identify transitions into pericentromeric regions, pericentromeres spanned ~54% of the genome and had an average recombination rate of 4,350 kb/cM versus 220 kb/cM in the euchromatic arms (**Supplementary Table 7**). The pericentromeres were primarily repetitive but, owing to their size, still contained 26.5% of the genes.

The majority of the repetitive elements in the genome were long terminal repeat (LTR) retrotransposons, and we identified 2,668 complete LTR retrotransposons and classified them into 165 families, including 65 *Ty1-copia*, 78 *Ty3-gypsy* and 22 unclassified families (**Supplementary Tables 8** and **9**). Although there were ancient elements that inserted into the genome more than 10 million years ago, ~75% (2,011/2,668) of the LTR retroelements integrated into *P. vulgaris* within the last 2 million years (**Supplementary Fig. 1**). Notably, the insertion times of 20% (543/2,668) of the elements were more recent than 0.5 million years ago—this is likely an underestimate, as our sequencing approach is biased against the annotation of completely identical LTRs. These results were similar to those in soybean[15] and suggest that LTR retrotransposons underwent recent amplification events in both legumes. The 165 LTR retrotransposon families varied in the copy number of complete elements: more than 78% (130/165) of the families had fewer than 10 complete retroelements, whereas 11 families had more than 50 complete elements and contained 63% (1,690/2,668) of the complete elements in the *P. vulgaris* genome. Some families showed extremely high copy numbers; for example, the pvRetroS2 family contained 446 complete elements (likely an underestimate, as some elements would not have been annotated uniquely).

We observed dense clusters of resistance-associated genes in the common bean genome. The majority of putative resistance-associated genes in plants encode nucleotide-binding and leucine-rich repeat domains and are collectively known as NB-LRR (NL) genes[15]. We identified 376 NL genes, of which 106 encoded an N-terminal Toll/interleukin-1 receptor (TIR)-like domain (TNLs) and 108 encoded an N-terminal coiled-coil domain (CNLs) (**Supplementary Table 10**). The majority of NL sequences were physically organized in complex clusters, often located at the ends of chromosomes (**Supplementary Fig. 16**). In particular, three large clusters were located at the ends of chromosomes Pv04, Pv10 and Pv11 and contained more than 40 NL genes that were enriched for CNL (Pv04 and Pv11) or TNL (Pv10) genes that colocalized with previously mapped genes related to disease resistance[16–21]. Local tandem duplications and ectopic recombination between clusters are involved in the evolution of these NL gene clusters[22].

Comparison of genome changes in sister legume species

P. vulgaris (common bean) and *G. max* (soybean) diverged ~19.2 million years ago but shared a whole-genome duplication (WGD) event ~56.5 million years ago[23]. *G. max* experienced an independent WGD ~10 million years ago[14]. These events were evident in plots of synonymous changes in coding sequences (*Ks*) between and within these genomes (**Supplementary Fig. 17**), which also showed that *P. vulgaris* has evolved more rapidly than *G. max* since they split from their last common ancestor. Assuming a divergence time of ~19.2 million years ago[23], the *Ks* value (synonymous substitution rate) for *P. vulgaris* was 1.4 times that of *G. max* (8.46×10^{-9} versus 5.85×10^{-9} substitutions/year).

We identified orthologous *P. vulgaris* and *G. max* genes using synteny and *Ks* values as criteria (**Supplementary Table 11**). Consistent with earlier work, there was extensive synteny between *P. vulgaris* and *G. max*, except in pericentromeric regions, where microcollinearity was often stretched out and thinned owing to genomic expansion in one or both genomes. Typically, two chromosomal blocks in *G. max* mapped to a single region of *P. vulgaris* owing to the most recent WGD in *G. max* (**Fig. 1**)[14,24,25]. Most of the *P. vulgaris* genes (91%; 24,861) were in identifiable synteny blocks in *G. max*, and 57% were in synteny blocks in *P. vulgaris* itself—a result of the ancient WGD event 55 million

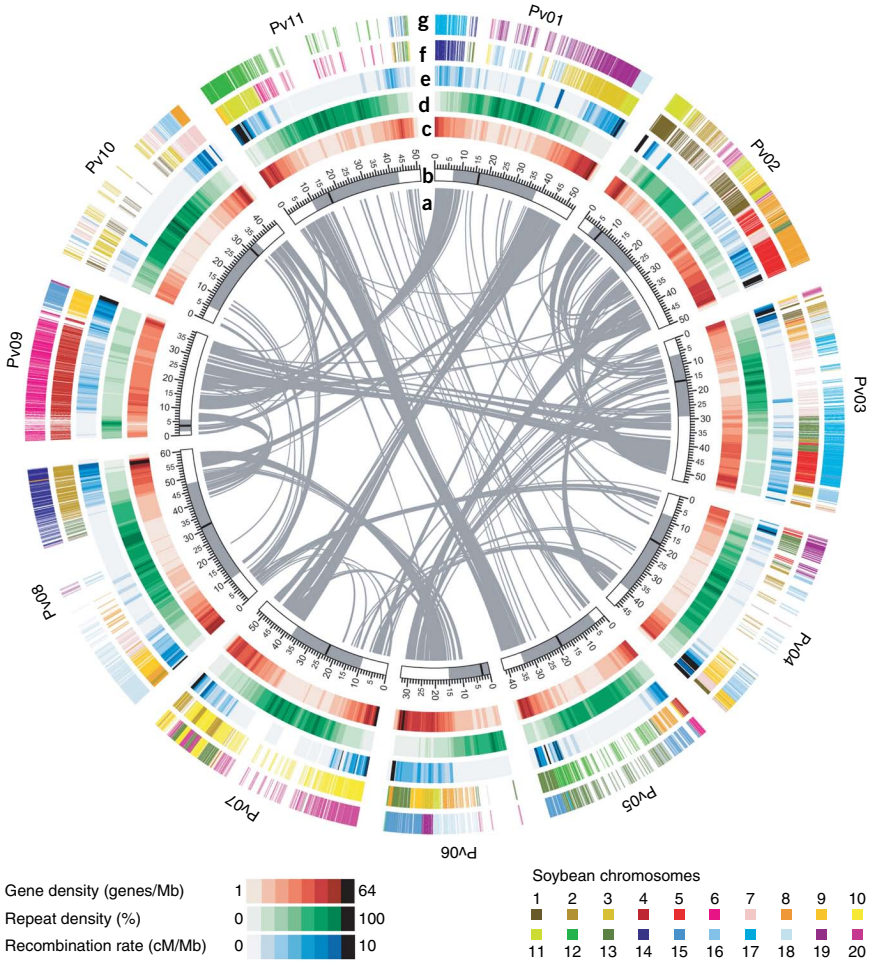


Figure 1 Structure of the *P. vulgaris* genome and synteny with the *G. max* genome. (**a**) Gray lines connect duplicated genes. (**b**) Chromosome structure with centromeric and pericentromeric regions in black and gray, respectively (scale is in Mb). (**c**) Gene density in sliding windows of 1 Mb at 200-kb intervals. (**d**) Repeat density in sliding windows of 1 Mb at 200-kb intervals. (**e**) Recombination rate based on the genetic and physical mapping of 6,945 SNPs and SSRs. (**f**,**g**) First syntenic region (**f**) and second *G. max* syntenic region (**g**) due to a lineage-specific duplication resulting in two chromosome segments for every segment in *P. vulgaris*.

years ago. Within synteny blocks, the *G. max*–*G. max* duplication had a mean of 33 genes/block, whereas the older, shared *P. vulgaris*–*G. max* WGD event had an average of 14 genes/block.

Evolution of gene pools in common bean

Mesoamerica has been suggested to be the center from which common bean originated, ultimately forming the distinct modern wild Andean and Mesoamerican gene pools[7]. To investigate the differentiation of these wild populations, we performed pooled resequencing of 30 individuals each from Mesoamerican and Andean wild populations (**Fig. 2** and **Supplementary Table 12**). Using π (the average pairwise nucleotide differences in a sample) and θ (the proportion of nucleotide polymorphisms in a sample), the Mesoamerican wild population (π (per bp) = 0.0061; θ (per bp) = 0.0041) was more diverse than the Andean wild population (π (per bp) = 0.0014; θ (per bp) = 0.0013). We used ~663,000 polymorphic sites (at least 5 kb from a gene and not in a repeat sequence) to estimate demographic parameters using the joint allele frequency spectrum (δaδi)[26] (**Supplementary Note**). The strong fixation index F_{ST} of ~0.34 between these two wild populations indicates that they have substantial allelic differentiation from each other. We estimated that divergence of the two wild pools occurred ~165,000 years ago, with an ancestral effective population size of 168,000. This date is earlier than a previous estimate of ~110,000 years ago but falls within the 95% confidence interval of the previous estimate, which was based on 13 loci from 24 wild genotypes[5], but it is later than other estimates of ~500,000 years ago[27]. The whole-genome analysis resulted in a much tighter confidence interval of 146,000–184,000 years ago.

Demographic inference for the wild Andean gene pool suggested that it was derived from the wild Mesoamerican population with a founding population of only a few thousand individuals (**Fig. 3a** and **Supplementary Note**). The wild Andean population showed no appreciable growth in effective population size for ~76,000 years after founding, although there was continual asymmetric gene flow between the two wild populations, with a higher Mesoamerican-to-Andean migration rate (**Supplementary Table 13**). The Andean population then underwent an exponential growth phase that began ~90,000 years ago and has continued to the present. The strong predomestication bottleneck in the Andean population has been observed in previous analyses[7,28,29]; in contrast, however, no detectable bottleneck was found for the wild Mesoamerican gene pool.

Domestication of common bean

To characterize diversity and differentiation within and between the Mesoamerican and Andean landraces (early domesticates), we sequenced 4 pooled populations representing distinct Mesoamerican landraces and 2 pooled populations representing distinct Andean landraces (n = 7–26 landraces). These landraces represent subpopulations from Mexico, Central America and South America with low levels of admixture (**Supplementary Fig. 18**). Because the four Mesoamerican and two Andean landrace populations are representative of the diversity of the original domestication populations, we combined SNP data from these populations to create a composite Mesoamerican and a composite Andean landrace SNP data set, respectively, for further analysis. This approach allowed us to distinguish selection from random fixation across the genome[30] and to search for signals associated with domestication events. The number of SNPs ranged from 8,890,318 for the wild Mesoamerican subpopulation to 1,397,405 SNPs for the Andean landrace subpopulation from Peru (**Supplementary Table 14**), and ~16% of these SNPs were within genes.

To characterize variation among the populations, we calculated diversity (π) and population differentiation (F_{ST}) statistics using data averaged over 10-kb windows with a 2-kb slide (10-kb/2-kb windows; **Supplementary Table 15**). Whereas the Mesoamerican landraces were less diverse than the wild Mesoamerican population, Andean landrace populations were more diverse than the wild Andean population, possibly owing to admixture with Mesoamerican populations and/or *de novo* mutation within the Andean gene pool. Diversity was further reduced within the Mesoamerican Central American and southern

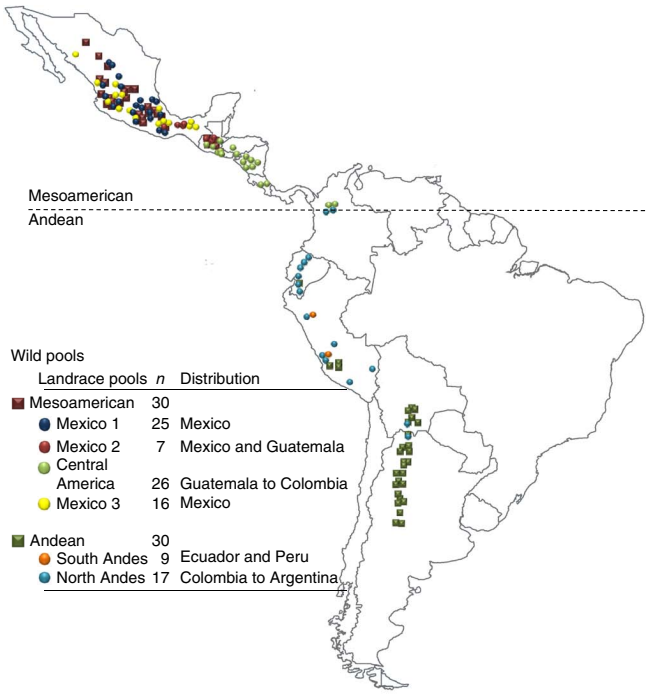


Figure 2 Geographic distribution of sampled genotypes.

Andean landraces, suggesting that these subpopulations underwent additional selection that might correspond to local adaptation.

Multiple results point to independent domestication events in the Mesoamerican and Andean gene pools, a feature observed for only a few modern crops. We characterized domestication of common bean at the genomic level by comparing wild and landrace populations across 10-kb/2-kb sliding windows, selecting windows that met strict composite criteria that required they be in the top 90% of the population's empirical distribution for both $\pi_{wild}/\pi_{landrace}$ ratios and F_{ST} values (**Figs. 3b**,**c** and **4**). We observed 930 windows in Mesoamerican populations (totaling 74 Mb of sequence) with both low diversity and high differentiation. Because low diversity and high differentiation are two features of selection[31], we consider these to be selection windows. Of these windows, 209 that were longer than 100 kb accounted for 70.1% of the total selection distance. Among the 750 selection windows in Andean populations exhibiting low diversity and high differentiation, 172 that were longer than 100 kb covered 69.8% of the total selection distance (60 Mb). As expected for independent Mesoamerican and Andean domestication events, these selection regions were distinct. Within the Mesoamerican landrace population, chromosomes Pv02, Pv07 and Pv09 accounted for 43% of the length (32.338 Mb), with 33.3% of chromosome Pv09 showing signatures of selection, whereas the Andean domestication event primarily involved chromosomes Pv01, Pv02 and Pv10 (**Fig. 4**). Interestingly, only 7.234 Mb of the regions predicted to be involved in domestication were shared by the two gene pools, suggesting different genetic routes to domestication.

We identified candidate genes associated with domestication using the same criteria applied to find selection windows (requiring that they be in the top 90% of the pool's empirical distribution for both $\pi_{wild}/\pi_{landrace}$ ratios and F_{ST} values). We identified 1,835 Mesoamerican and 748 Andean candidate genes associated with domestication (**Supplementary Tables 16** and **17**), and all candidates had a negative Tajima's *D* value, indicating positive selection. Most notably, only 59 of the candidate genes (3% of the Mesoamerican and 8% of the Andean candidates) were shared by the 2 landrace populations. For the 59 common candidates, the mean F_{ST} value was 0.67, suggesting selection on different alleles or the appearance of unique mutations in the two gene pools. This finding is consistent with evidence at the *PvTFL1y* determinancy locus that was independently derived in each gene pool[32] but contrasts with evidence in rice, where a domestication locus appeared uniquely in one gene pool, *indica* or *japonica*, and was transferred to the other pools[33]. Most Mesoamerican candidate genes (n = 1,561; 85%) were located in 10-kb selection windows, whereas only 48.1% of the Andean candidate genes were within such windows (**Supplementary Table 18**). The effects of domestication were uneven across the Mesoamerican subpopulations: we detected only 418 candidates in the Mesoamerican Central American landrace population compared to 1,424 candidates

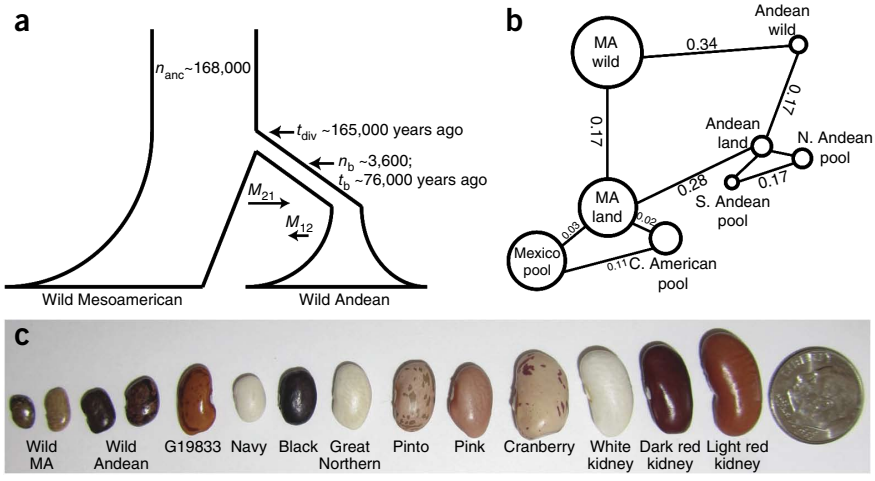


Figure 3 Evolution and domestication of common bean. (**a**) Divergence of the wild Mesoamerican and Andean common bean pools. The wild Andean gene pool diverged from the wild Mesoamerican gene pool ~165,000 years ago, with a small founding population and a strong bottleneck that lasted ~76,000 years. The bottleneck was followed by an exponential growth phase extending to the present day. Asymmetric gene flow between the two pools had a key role in maintaining genetic diversity, especially in the Andean population, with average migration rates M_{21} = 0.135 (wild Mesoamerican to wild Andean) and M_{12} = 0.087 (wild Andean to wild Mesoamerican). This scenario conforms to the Mesoamerican origin model of the common bean, with an Andean bottleneck that predated domestication. (n_{anc}, size of ancestral population; t_{div}, start of bottleneck; n_b, size of bottleneck population; t_b, length of bottleneck) (**b**) Population genomic analysis based on SNP data from the resequencing of DNA pools for common bean. The size of the circle for each pool is proportional to the π value for the pool. For a reference, π = 0.0061 for the wild Mesoamerican (MA) pool. F_{ST} statistics, representing the differentiation of any two pools, are noted on the lines (not proportional) connecting pools. Data are average statistics across all 10-kb/2-kb sliding/discarding windows with <50% called bases. Land, landrace; N, north; S, south; C, central. (**c**) Variation in seed size in common bean. The seeds of wild Mesoamerican and Andean beans (two each) are smaller than the seeds corresponding to the reference genotype (G19833) and the multiple market classes of common beans grown in the United States (navy to light red kidney).

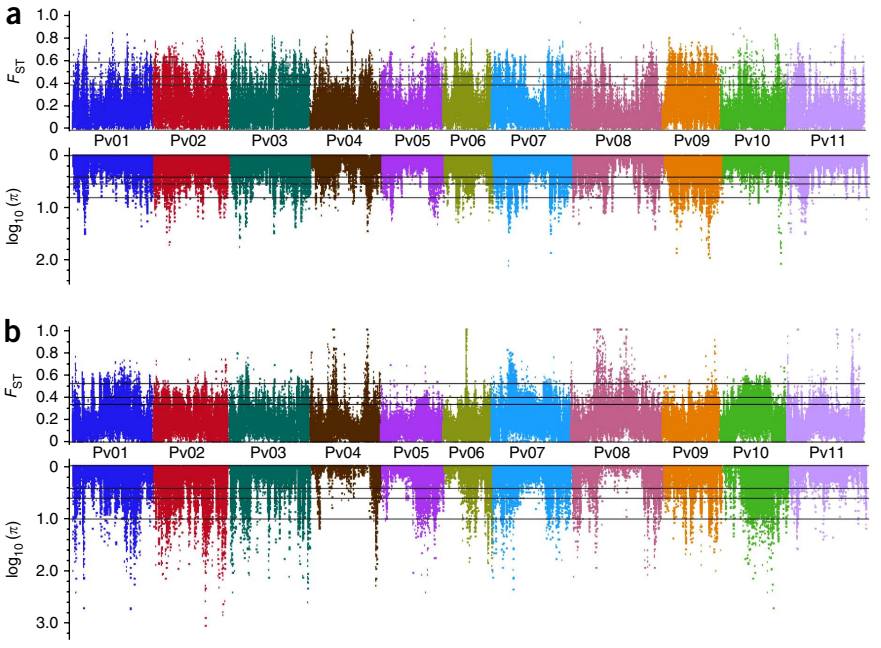


Figure 4 Differentiation and reduction in diversity during the domestication of common bean. (**a**,**b**) Genome-wide view in 10-kb/2-kb sliding windows of differentiation (F_{ST}) and reduction in diversity (π ratio) statistics associated with domestication within the common bean Mesoamerican (**a**) and Andean (**b**) gene pools. Log_{10} π ratios less than zero are not shown. Lines represent the 90%, 95% and 99% tails for the empirical distribution of each statistic.

in the Mesoamerican Mexican landraces. The fact that only 33 of these genes were shared by these 2 subpopulations indicates unique evolutionary trajectories among subpopulations of the Mesoamerican gene pool. Within the Andean gene pool, none of the candidate genes from the northern and southern Andean landrace populations were shared. These results demonstrate that the sexually compatible Mesoamerican and Andean lineages with similar morphologies and life cycles underwent independent selection upon distinct sets of genes. This is in contrast to the situation in rice, where many major domestication genes were shared by gene flow between the *indica* and *japonica* types[34].

Domestication had distinct effects on genes involved in flowering[35] in the two gene pools. Whereas the principal floral integrator genes *SOC1* and *FT*[35] were not candidate domestication genes in either pool, 25 Mesoamerican and 13 Andean genes that are in pathways that control these 2 genes were candidate genes for domestication. For example, within the vernalization pathway, orthologs of *VRN1* (*Phvul.003G033400*) and *VRN2* (*Phvul.002G000500*) were Mesoamerican candidate genes, and orthologs of *FRL1* (*Phvul.006G053200*) and *TFL2* (*Phvul.009G117500*) were Andean candidate genes. *COP1* encodes a photoperiod pathway regulator that controls *FT* through *CO*. The Mesoamerican ortholog of *COP1* was a candidate domestication gene, and *Phvul.006G165300*, a *CUL4* ortholog that encodes a protein that is part of a complex that along with COP1 regulates *CO*[36], was an Andean candidate gene for domestication. This finding demonstrates independent selection on genes encoding different members of the same protein complex. The only shared domestication candidates were *Phvul.007065600*, an ortholog of *AGL42*, which regulates flowering through the gibberellin pathway, and *Phvul.009G203400*, an ortholog of *FUL*, which regulates *SOC1*.

Increased plant size is typically associated with plant domestication[37], and multiple Mesoamerican candidate genes influence this trait. *Phvul.011G213300* is an ortholog of *Arabidopsis thaliana BB*, a component of the ubiquitin ligase degradation pathway that controls flower and stem size[38], and *Phvul.009G040200* is an ortholog of *BIN4*, which regulates cell expansion and final plant size[39]. Multiple candidate genes for domestication were also components of nitrogen metabolism pathways, which directly affect plant size. The Mesoamerican candidate gene *Phvul.008G168000* encodes nitrate reductase, a critical element for plant and seed growth, which genetically maps to the *SW8.2* quantitative trait locus (QTL) for seed weight[40]. Other candidate genes for domestication involved in nitrogen metabolism included the Mesoamerican (*Phvul.005G132200*) and Andean (*Phvul.002G242900*) nitrogen transporters and the Mesoamerican asparagine synthase (*Phvul.006G069300*).

Increased seed size is a major phenotypic shift associated with the domestication of the common bean[41] and other legumes[42] and

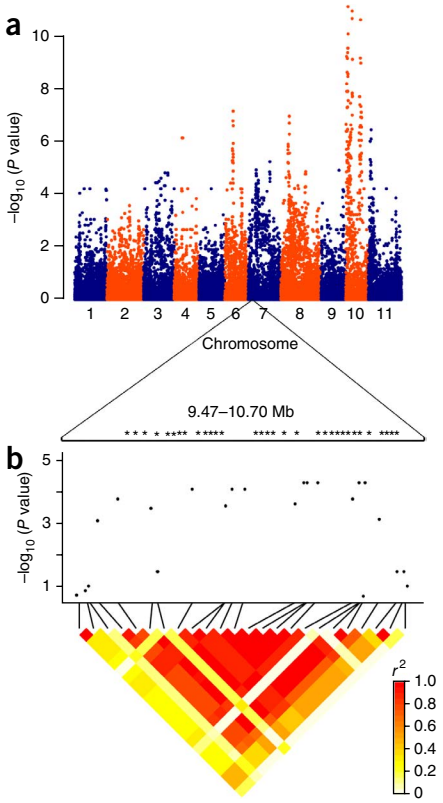


Figure 5 Genome-wide association analysis of seed weight. (**a**) A 280-member panel of Mesoamerican cultivars was grown in 4 locations in the United States. Phenotypic data were coupled with 34,799 SNP markers and analyzed using a mixed-model analysis that controlled for population structure and genotype relatedness. (**b**) A close-up view of the GWAS results for seed weight and linkage disequilibrium (r^2) around a 1.23-Mb Mesoamerican sweep window on Pv07. The positions of candidate genes for domestication are noted by asterisks above the GWAS display. The candidates range from *Phvul.007G094299* to *Phvul.007G.99700* (**Supplementary Note**).

distinguishes the many types of beans that humans consume. We surveyed the Mesoamerican domestication candidates for genes previously shown to be associated with seed weight[43] and used the whole-genome sequence for a genome-wide association study (GWAS; **Fig. 5a**) to understand the genetic architecture of seed weight in modern Mesoamerican cultivars. We found 15 candidate genes previously shown to be involved in seed weight (**Supplementary Table 19**). Among these are nearly all the components of the cytokinin synthesis and multiple-component phosphorelay regulatory system (**Supplementary Fig. 19**). Included are *Phvul.002G082400*, which encodes a protein that transmits the phosphosignal in response to regulators, and three type B response regulator transcription factors (*Phvul.003G017000*, *Phvul.003G110100* and *Phvul.009G088900*), which in turn activate a number of downstream genes[44]. An additional candidate gene, *Phvul.01G038800*, has orthologs that encode cytokinin oxidase/dehydrogenase proteins, which regulate the pathway by degrading active cytokinin. The relevance of these genes as candidate loci associated with seed weight is supported by work in *Arabidopsis*, where orthologs of the candidate genes in the cytokinin pathway have been shown in transgenic studies to regulate seed size and/or weight[43]. In contrast, however, none of these genes were Andean domestication candidates.

GWAS analysis for seed weight confirmed three of these domestication candidates. It was not possible to confirm the other 12 candidates by GWAS because Mesoamerican domestication reduced diversity to near homozygosity, such that associations could not be found (**Supplementary Table 20**). GWAS analysis was able to place 75 domestication candidate genes within 50 kb of a SNP significantly ($P < 1.0 \times 10^{-4}$) associated with seed weight, and a significantly associated SNP was found within eight candidate genes (**Supplementary Table 21**). One sweep window on Pv07 (9.662–10.662 Mb) contained 33 domestication candidates and was located in a GWAS peak that exhibited extensive linkage disequilibrium (**Fig. 5b**). By GWAS, we also detected candidate genes for seed weight that resulted from modern breeding of the common bean. These included 15 improvement-related genes previously shown to be associated with seed weight, 5 of which function in the cytokinin regulation/degradation pathway (**Supplementary Table 22**). Finally, three genes in complete linkage disequilibrium with equally significant association ($P = 6.3 \times 10^{-6}$) were located in a Pv07 QTL for seed weight that has been replicated in many experiments[45].

DISCUSSION

Common bean is the most important grain legume for human consumption and is an especially nutrient-dense food in developing parts of the world. Improvement of common bean will require a more fundamental understanding of the genetic basis of how it responds to biotic and abiotic stresses. The clustering of resistance-associated genes in a few genomic locations suggests that stacking resistances between clusters should be relatively easy but that stacking multiple resistance genes located within a single physical cluster and then combining these traits by breeding may prove more challenging. The observation that the dual domestication events for common bean had few selective sweeps in common leads us to posit that domestication, previously thought to typically be associated with selection at a few major loci, can also be achieved via multiple genetic pathways resulting in similar or the same phenotypes (for example, seed size). In addition, the lack of correspondence between selective sweeps in domestication and genetic bottlenecks imposed by breeding indicates that domestication-derived traits were fixed early and that subsequent selection was likely on traits for local adaptation and desired seed and plant traits. Together, these findings provide information on regions of the genome that have undergone intense selection, either during domestication or early improvement, and thus provide targets for future crop improvement efforts, as valuable alleles will have been lost during early selection.

URLs. Food and Agricultural Organization of the United Nations (FAO) statistics, http://faostat.fao.org/site/291/default.aspx; Plant DNA C-values Database, http://www.kew.org/cvalues/; Phytozome transposon database, http://www.Phytozome.net/; RepeatMasker, http://www.repeatmasker.org/; MEGA 4, http://www.megasoftware.net/mega4/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Assembly and annotation are available at http://www.phytozome.net/commonbean.php and have been deposited in GenBank under accession ANNZ01000000.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract DE-AC02-05CH11231. This research was funded by grants from the US Department of Agriculture–National Institute for Food and Agriculture (2006-35300-17266) and the National Science Foundation (DBI 0822258) to S.A.J. and from the US Department of Agriculture–Cooperative State Research, Education and Extension Service (2009-01860 and 2009-01929) to S.A.J. and P.E.M., respectively.

AUTHOR CONTRIBUTIONS

J.S., P.E.M., D.S.R. and S.A.J. conceived the study and jointly wrote the manuscript with S.B.C. Genomic clones and DNA were provided by R.A.W., Y.Y., D.K., R.L. and M.B. The following analyses were performed by the indicated authors: repeat annotation, D.G.; identification of resistance genes, V.G., M.M.S.R. and V.T.; genetic mapping, P.B.C., Q.S., J.R., D.L.H. and G.J.; sequencing, assembly and/or annotation, J.G., J.J., S.S., K.B., M.C., D.M.G., U.H., M.W. and M.Z.; comparative, population and/or evolutionary analyses, S.M., G.A.W., S.B.C., C.C., S.M.M., B.A., M.T.-T. and M.G.; and GWAS, S.M.M., M.A.B., P.G., J.D.K., P.N.M., J.M.O. and C.A.U.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Anderson, J.W. *et al.* Hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men. *Am. J. Clin. Nutr.* **40**, 1146–1155 (1984).
2. Geil, P. & Anderson, J. Nutrition and health implications of dry beans: a review. *J. Am. Coll. Nutr.* **13**, 549–558 (1994).
3. Cichy, K.A., Caldas, G.V., Snapp, S.S. & Blair, M.W. QTL analysis of seed iron, zinc, and phosphorus levels in an Andean bean population. *Crop Sci.* **49**, 1742–1750 (2009).
4. Beebe, S. Common bean breeding in the tropics. *Plant Breed. Rev.* **36**, 357–426 (2012).
5. Mamidi, S. *et al.* Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. *Heredity* **110**, 267–276 (2013).
6. Bitocchi, E. *et al.* Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytol.* **197**, 300–313 (2013).
7. Bitocchi, E. *et al.* Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proc. Natl. Acad. Sci. USA* **109**, E788–E796 (2012).

8. Gepts, P., Osborn, T., Rashka, K. & Bliss, F. Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): evidence for multiple centers of domestication. *Econ. Bot.* **40**, 451–468 (1986).
9. Mamidi, S. *et al.* Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Funct. Plant Biol.* **38**, 953–967 (2011).
10. Zizumbo-Villarreal, D. & Colunga-GarcíaMarín, P. Origin of agriculture and plant domestication in West Mesoamerica. *Genet. Resour. Crop Evol.* **57**, 813–825 (2010).
11. Singh, S.P., Gepts, P. & Debouck, D.G. Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot.* **45**, 379–396 (1991).
12. McClean, P.E., Lee, R., Otto, C., Gepts, P. & Bassett, M. Molecular and phenotypic mapping of genes controlling seed coat pattern and color in common bean (*Phaseolus vulgaris* L.). *J. Hered.* **93**, 148–152 (2002).
13. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
14. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
15. Meyers, B.C., Kaushik, S. & Nandety, R.S. Evolving disease resistance genes. *Curr. Opin. Plant Biol.* **8**, 129–134 (2005).
16. Geffroy, V. *et al.* Molecular analysis of a large subtelomeric nucleotide-binding-site–leucine-rich-repeat family in two representative genotypes of the major gene pools of *Phaseolus vulgaris*. *Genetics* **181**, 405–419 (2009).
17. Geffroy, V. *et al.* Identification of an ancestral resistance gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen *Colletotrichum lindemuthianum*. *Mol. Plant Microbe Interact.* **12**, 774–784 (1999).
18. Innes, R.W. *et al.* Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiol.* **148**, 1740–1759 (2008).
19. Chen, N.W.G. *et al.* Specific resistances against *Pseudomonas syringae* effectors AvrB and AvrRpm1 have evolved differently in common bean (*Phaseolus vulgaris*), soybean (*Glycine max*), and *Arabidopsis thaliana*. *New Phytol.* **187**, 941–956 (2010).
20. Geffroy, V. *et al.* A family of LRR sequences in the vicinity of the *Co-2* locus for anthracnose resistance in *Phaseolus vulgaris* and its potential use in marker-assisted selection. *Theor. Appl. Genet.* **96**, 494–502 (1998).
21. Miklas, P.N., Kelly, J.D., Beebe, S.E. & Blair, M.W. Common bean breeding for resistance against biotic and abiotic stresses: from classical to MAS breeding. *Euphytica* **147**, 105–131 (2006).
22. David, P. *et al.* A nomadic subtelomeric disease resistance gene cluster in common bean. *Plant Physiol.* **151**, 1048–1065 (2009).
23. Lavin, M., Herendeen, P.S. & Wojciechowski, M.F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst. Biol.* **54**, 575–594 (2005).
24. Gill, N. *et al.* Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* **151**, 1167–1174 (2009).
25. McClean, P.E., Mamidi, S., McConnell, M., Chikara, S. & Lee, R. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC Genomics* **11**, 184 (2010).
26. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
27. Chacón S, M.I., Pickersgill, B. & Debouck, D.G. Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theor. Appl. Genet.* **110**, 432–444 (2005).
28. Kwak, M. & Gepts, P. Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). *Theor. Appl. Genet.* **118**, 979–992 (2009).
29. Rossi, M. *et al.* Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. *Evol. Appl.* **2**, 504–522 (2009).
30. Rubin, C.-J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).
31. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
32. Repinski, S.L., Kwak, M. & Gepts, P. The common bean growth habit gene *PvTFL1y* is a functional homolog of *Arabidopsis TFL1*. *Theor. App. Genet.* **124**, 1539–1547 (2012).
33. Sweeney, M.T. *et al.* Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genet.* **3**, e133 (2007).
34. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
35. Fornara, F., de Montaigu, A. & Coupland, G. SnapShot: control of flowering in *Arabidopsis thaliana*. *Cell* **141**, 550 (2010).
36. Chen, H. *et al. Arabidopsis* CULLIN4-damaged DNA binding protein 1 interacts with CONSTITUTIVELY PHOTOMORPHOGENIC1–SUPPRESSOR OF PHYA complexes to regulate photomorphogenesis and flowering time. *Plant Cell* **22**, 108–123 (2010).
37. Gepts, P. Crop domestication as a long-term selection experiment. *Plant Breed. Rev.* **24**, 1–44 (2004).
38. Disch, S. *et al.* The E3 ubiquitin ligase BIG BROTHER controls *Arabidopsis* organ size in a dosage-dependent manner. *Curr. Biol.* **16**, 272–279 (2006).
39. Breuer, C. *et al.* BIN4, a novel component of the plant DNA topoisomerase VI complex, is required for endoreduplication in *Arabidopsis*. *Plant Cell* **19**, 3655–3668 (2007).
40. Pérez-Vega, E. *et al.* Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **120**, 1367–1380 (2010).
41. Koinange, E.M., Singh, S.P. & Gepts, P. Genetic control of the domestication syndrome in common bean. *Crop Sci.* **36**, 1037–1045 (1996).
42. Weeden, N.F. Genetic changes accompanying the domestication of *Pisum sativum*: is there a common genetic basis to the 'domestication syndrome'for legumes? *Ann. Bot.* **100**, 1017–1025 (2007).
43. Van Daele, I. *et al.* A comparative study of seed yield parameters in *Arabidopsis thaliana* mutants and transgenics. *Plant Biotechnol. J.* **10**, 488–500 (2012).
44. Hwang, I., Sheen, J. & Muller, B. Cytokinin signaling networks. *Annu. Rev. Plant Biol.* **63**, 353–380 (2012).
45. González, A.M., De la Fuente, M., De Ron, A.M. & Santalla, M. Protein markers and seed size variation in common bean segregating populations. *Mol. Breed.* **25**, 723–740 (2010).

ONLINE METHODS

Sequencing. The majority of *de novo* genome sequencing reads were collected with standard sequencing protocols provided by the manufacturer on Roche 454 XLR and Illumina HiSeq 2000 machines at the Department of Energy Joint Genome Institute in Walnut Creek, California. Two types of linear 454 data were collected, standard XLR data (31 runs; 10.7 Gb) and FLX+ data (8.5 runs; 5.615 Gb). Six different paired 454 libraries were created, three libraries with average insert sizes of 2.8–4.8 kb, 1 library with average insert size of 8.0 kb, 1 library with average insert size of 9.2 kb, 1 library with average insert size of 11.9 kb and 1 library with average insert size of 12.2 kb, and were sequenced by standard XLR (26.5 runs; 6.282 Gb of useable data). Two standard 400-bp fragment libraries were sequenced at 2 × 101 bp (four channels; 135.8 Gb) on an Illumina HiSeq 2000. Two fosmid libraries (328,704 reads; 223.9 Mb) with 35.0-kb and 36.0-kb insert sizes and 3 BAC libraries (89,017 reads; 55.1 Mb) with 127.0-kb, (92,160 reads; 65.9 Mb), 135.3-kb (81,408 reads; 57.6 Mb) and 122.0-kb average insert sizes were sequenced on both ends with Sanger sequencing for a total of 591,289 Sanger reads of 402.5 Mb of high-quality sequence. Fosmid-end and BAC-end sequence data were collected using standard protocols at the HudsonAlpha Institute in Huntsville, Alabama, and at the Arizona Genomics Institute in Tucson, Arizona. Sixty *P. vulgaris* genotypes representing 30 wild Mesoamerican and 30 wild Andean individuals were pooled into 2 sequencing libraries, and 54× and 4.9× genome equivalents were collected on a HiSeq 2000 with unamplified libraries. Similarly, 100 genotypes from 6 individual landrace classes, selected from a structure analysis, were pooled into 6 libraries, and sequencing depths from 3.4 to 7.1× were achieved.

Construction of the genetic map. We obtained 19,619 Mb of 121-bp paired-end Illumina Genome Analyzer IIx short reads from a diverse set of genotypes for common bean. Reads were aligned to the genome reference sequences for common bean with 14× coverage, and SNPs were called using CASAVA1.7 software (Illumina, 2010) with the default settings. After filtering out A/T or G/C SNPs, SNPs with Ns in the 60 nt of flanking sequence and SNPs residing within 25 nt of another SNP, a total of 992,682 SNPs remained. Using these SNPs, an Illumina Infinium BeadChip (BARCBEAN6K_1 with 5,232 SNPs) was designed. The SNPs for BARCBEAN6K_1 were selected to optimize polymorphism among the various common bean market classes, and, when possible, SNPs were targeted to sequence scaffolds (>10 kb) in an early *P. vulgaris* assembly. A mapping population of 267 F_2 progeny from a cross of the common bean cultivars Stampede and Red Hawk developed at North Dakota State University was genotyped with the BARCBEAN6K_1 BeadChip. An additional BeadChip (BARCBEAN6K_2 with 5,514 SNPs) was designed using the same steps as with the *P. vulgaris* v0.9 assembly, with markers selected to anchor and orient additional scaffold sequences and used to type the same population. Both BeadChips and 261 SSR markers were also used to genotype 88 F_5-derived RILs from the cross of the Stampede and Red Hawk cultivars. SSRs were selected from sequence scaffolds in the *P. vulgaris* 8× assembly, PCR markers were designed and fragment length polymorphisms were assessed as described in Song *et al.*[46]. Linkage maps were constructed using JoinMap 4.0 (ref. 47) software on the basis of the 6,531 polymorphic SNPs from these 2 BeadChips and 484 SNP loci that were genotyped with the Illumina GoldenGate assay at the US Department of Agriculture–Agricultural Research Service in Beltsville, Maryland[48], as well as 261 SSR markers and 25 framework markers. The final map contained 7,276 SSR and SNP markers arranged in 11 linkage groups via framework markers.

Genome assembly and construction of pseudomolecule chromosomes. Before assembly, reads corresponding to organelle DNA were removed by screening against identified fragments of mitochondria, chloroplast and rDNA. For Roche 454 linear reads, any read <200 bp in length was discarded. Roche 454 paired reads were split into pairs, and any pair with a read shorter than 50 bp was discarded. An additional deduplication step was applied to the 454 paired libraries that identified and retained only one copy of each PCR duplicate. All remaining 454 reads were compared against 24.1 Gb of trimmed HiSeq 2000 V3 reads from two separate libraries, and any insertion-deletions in the 454 reads were corrected to match the Illumina alignments. Before assembly, 454 reads that contained >80% 24-mers that occurred ≥400 times in the data set were removed to reduce improper assembly of transposon sequences. Sequence reads were assembled using our modified version of Arachne v.20071016 (ref. 49) with parameters maxcliq1 = 250 and BINGE_AND_PURGE = True, bless = False BINGE_AND_PURGE = True lap_ratio = 0.8 max_bad_look = 2000 (note: Arachne error correction was on). An additional filtering step to remove contigs of <300 bp in length or with fewer than four reads was applied. This produced 1,627 scaffold sequences, with a scaffold L50 value of 6.0 Mb; 171 scaffolds were greater than 100 kb in length, and the total genome size was 474.3 Mb (**Supplementary Table 2**). Scaffolds were screened against bacterial proteins, organelle sequences and the GenBank nr database and were removed if found to be a contaminant. Additional scaffolds were removed if they (i) consisted of >95% 24-mers that occurred four other times in scaffolds greater than 50 kb in length, (ii) contained only unanchored RNA sequences or (iii) were less than 1 kb in length.

The 7,015 markers from the genetic map were aligned to the assembly using BLAT[50] (parameters: *-t* = dna -q = dna -minScore = 200 –extendThroughN). Positions of SSR markers were determined using E-PCR[51]. Scaffolds were broken if they contained linkage group or syntenic discontiguity coincident with an area of low BAC or fosmid coverage. A total of 71 breaks were executed and 284 joins were made to form the final assembly consisting of 11 pseudomolecule chromosomes. Each chromosome join was padded with 10,000 Ns to indicate unsized map joins. The final assembly contained 708 scaffolds (41,391 contigs) that cover 472.5 Mb of the genome with a contig N50 value of 39.5 kb and a scaffold N50 value of 50.4 Mb.

Completeness of the euchromatic portion of the genome assembly was assessed using 108,874 *P. vulgaris* EST sequences obtained from GenBank. These sequences were aligned to the assembly to estimate completeness using BLAT (parameters: *-t* = dna -q = rna –extendThroughN). Alignments that comprised ≥90% base-pair identity and ≥85% EST coverage were retained. The screened alignments indicated that 102,254 of the 108,874 cDNAs (93.92%) aligned to the assembly. At least 30% of the ESTs that did not align were bacterial or fungal contaminants. In addition, BAC clones from euchromatic regions and moderately to highly repetitive regions were sequenced and compared to the assembly (**Supplementary Figs. 19–23**).

Annotation. We constructed 43,627 transcript assemblies from about 727 million reads of paired-end Illumina RNA-seq data. These transcript assemblies were constructed using PERTRAN (S.S., unpublished data). We built 47,464 transcript assemblies using PASA[52] from 79,630 *P. vulgaris* Sanger ESTs and the RNA-seq transcript assemblies. Loci were identified by transcript assembly alignments and/or EXONERATE alignments of peptides from *Arabidopsis*, poplar, *Medicago truncatula,* grape (*Vitis vinifera*) and rice (*Oryza sativa*) peptides to the repeat-soft-masked genome using RepeatMasker[53] on the basis of a transposon database developed as part of this project (see URLs) with up to 2,000-bp extension on both ends, unless they extended into another locus on the same strand. Gene models were predicted by the homology-based predictors FGENESH+ (ref. 53), FGENESH_EST (similar to FGENESH+; EST as splice-site and intron input instead of peptide/translated ORF) and GenomeScan[54]. The highest scoring predictions for each locus were selected using multiple positive factors, including EST and peptide support, and one negative factor—overlap with repeats. Selected gene predictions were improved by PASA, including by adding UTRs, correcting splicing and adding alternative transcripts. PASA-improved gene model peptides were subjected to peptide homology analysis with the above-mentioned proteomes to obtain Cscore values and peptide coverage. Cscore is the ratio of the peptide BLASTP score to the mutual best hit BLASTP score, and peptide coverage is the highest percentage of peptide aligned to the best homolog. A transcript was selected if its Cscore value was greater than or equal to 0.5 and its peptide coverage was greater than or equal to 0.5 or if it had EST coverage but the proportion of its coding sequence overlapping repeats was less than 20%. For gene models where greater than 20% of the coding sequence overlapped with repeats, the Cscore value was required to be at least 0.9 and homology coverage was required to be at least 70% to be selected. Selected gene models were subjected to Pfam analysis, and gene models whose encoded peptide contained more than 30% Pfam transposon element domains were removed. The final gene set consisted of 27,197 protein-coding genes and 31,638 protein-coding transcripts.

doi:10.1038/ng.3008

Repeat analysis. In addition to the genome sequence, 15 publicly available BAC sequences for common bean were also downloaded from GenBank for a total of 2.2 Mb of sequence, including from accessions DQ205649, DQ323045, FJ817289–FJ817291 and GU215957–GU215966. Transposon annotation was conducted using different methods according to the sequence structures and transposases of various transposons. To annotate LTR retrotransposons, the genome sequence was screened with LTR_Finder[35] using default parameters, except that we set a 50-bp minimum LTR length and 50-bp minimum distance between LTRs. All predicted LTR retrotransposons were manually inspected to eliminate incorrectly predicted sequences, including tandem repeats, nested transposons, incomplete DNA transposons and other sequences. The internal sequences of LTR retrotransposons were used to perform BLASTX and/or BLASTP searches to define superfamilies: *Ty1-copia*, *Ty3-gypsy* or other. LINEs (long interspersed elements) were predicted on the basis of the non-LTR retrotransposase and polyA sequences. SINEs (short interspersed elements) were annotated with the polyA structure feature and combined with BLAST searches. To find DNA transposons, conserved domains for transposases from different reported superfamilies were used as queries to search the common bean genome. The matching sequences and flanking sequence (10 kb on each side) were extracted to conduct BLASTN searches to identify complete DNA transposons by terminal inverted repeats (TIRs) and target size duplication (TSD). Furthermore, MITEs-Hunter software[36] was also used to identify DNA elements. The annotated transposons and two reported LTR retrotransposons, pva1-118d24-re-5 (FJ402927) and Tpv2-6 (AJ005762), were combined and used as a transposon library to screen the genome using RepeatMasker with default settings except that we used the 'nolow' option to avoid masking low-complexity DNA or simple repeats. Transposons were summarized according to names, subclasses and classes, and overlapping regions in the RepeatMasker output file were counted once (**Supplementary Table 9**).

To estimate the insertion times of LTR retrotransposons, the 5′ and 3′ LTRs for each full-length LTR retroelement were aligned and used to calculate the nucleotide divergence rate with the Kimura-2 parameter using MEGA 4. The insertion date (T) was estimated with the formula $T = K/2r$, where K is the average number of substitutions per aligned site and r is an average substitution rate. We used the average substitution rate of 1.3×10^{-8} substitutions per synonymous site per year[55] to calibrate the insertion times.

Identification of disease resistance genes. NL proteins were identified in an iterative process. First, an HMM (Hidden Markov model) search of the predicted protein sequences identified sequences containing the NB-ARC domain. The 'trusted cutoff' of the NB-ARC domain HMM (PF00931) established by Pfam[56] was used as the threshold for detecting NBS domains. We identified 398 predicted proteins corresponding to 342 annotated genes that encoded homologs of NL proteins. To identify diverse homologs, all the NL predicted protein sequences were used as queries for TBLASTN[57] against the entire genome. All resulting sequences (E value < 1×10^{-10}) were manually inspected using Artemis[58]. This procedure identified an additional 38 putative NL genes that were not part of the genome annotation. A new identifier was created for each missing gene (with last digits set as 50). NL genes were assessed manually in Artemis software for the presence of sequences encoding TIR (PF01582), NB-ARC (PF00931) and LRR (PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13504 and PF13855) domains with HMMer using the trusted cutoffs defined in Pfam. Coiled-coil domains were identified using Coils[59] with a 14-amino-acid search window and a cutoff score of 2.9. Artemis was used for further manual analysis. Gene models with stop codons and/or frameshifts were classified as pseudogenes.

Development of wild and landrace pools for sequencing of common bean. Initially, 126 wild and 179 landrace genotypes, collected from the full geographic range of the species, were scored with 22 indel markers distributed throughout the genome. A Bayesian analysis was performed on the genotype data within each of the two groups using STRUCTURE software[60,61] with the parameters outlined previously[62]. For the wild genotypes where k is the number of populations, $k = 2$ best fit the data[63], and, for the landraces, $k = 6$ defined 3 Mexican subpopulations, 1 Central American subpopulations and 2 Andean subpopulations. A genotype was assigned to a subpopulation if its subpopulation parentage was >70%. DNA pools for resequencing were created by selecting individuals with high subpopulation membership (>98% for wild subpopulations and >90% for landrace subpopulations; **Supplementary Fig. 18**). In adopting other approaches[30,31], several individual-pool SNP data were combined with other pool SNP data to create a pool SNP data set representing a putative ancestral state.

Pooled DNA sequencing and SNP identification. DNA from each of these pools was sequenced to ~4× depth using Illumina technology (**Supplementary Table 12**). Each read was mapped to the v1.0 version of the assembled reference genome using Burrows-Wheeler Aligner (BWA)[64] with the maximum edit distance set to 8. All reads with a mapping quality score of less than 25 were discarded. An mpileup file was created for each sequenced pool using SAMtools[65] with the –BA options. VarScan 2.2.10 (ref. 66) used the mpileup file for SNP calling with the following parameters: minimum coverage = 5, minimum consensus quality = 25 and minimum variant frequency = 0.01. To further reduce SNP call quality, SNPs were discarded (i) if the reference or variant allele was an N; (ii) if more than one variant allele was observed; and (iii) if the variant allele was a single-nucleotide indel. The minimum number of reads required for the reference or variant allele was three. The number of SNPs ranged from 8,890,318 for the wild Mesoamerican pool to 1,397,405 for the Peru landrace pool (**Supplementary Table 14**). Among wild genotypes, 10,158,326 SNPs were observed, whereas the Mesoamerican landrace genotypes contained 9,661,807 SNPs and the Andean landrace genotypes contained 3,154,648 SNPs. For individual and combined pools, the proportion of SNPs found within genes was ~16%, indicating that genes were not disproportionately prone to more (or less) variation.

Demographic modeling. To minimize bias in demographic inferences due to selection, we used neutral sites defined to be at least 5 kb away from a gene (as annotated in the gff3 file v1.0) and not located in repetitive regions. The number of different haplotypes for each pooled sample was close to 30. Data were thus down-sampled to 25 haplotypes for each pool via hypergeometric projection (random sampling of 25 alleles without replacement), from which the joint allele frequency spectrum (jAFS) was derived. To eliminate spurious singletons, we excluded sites appearing as singletons in either of the two pools, resulting in a total of 663,000 polymorphic sites for jAFS.

We compared different demographic models on the basis of the relative log likelihoods of the models given the observed site frequency spectrum. Asymmetric migration rates were assumed in the model (**Fig. 1**). To infer model parameters, we ran δaδi simulations with different starting points in an eight-dimensional parameter space until convergence was achieved. Parameter values for the best-fit model are listed in **Supplementary Table 13**, using a base substitution rate $\mu = 8.46 \times 10^{-9}$ substitutions/bp/year (S.B.C., unpublished data) derived from silent sites. To estimate parameter uncertainties, we divided the genome into 10-cM segments and performed 100 bootstraps on the chromosome segments. Confidence intervals were derived on the basis of simulation results for the bootstrapped samples (**Supplementary Table 13**) as were comparisons between model prediction and observed data (**Supplementary Figs. 24** and **25**).

Population genetics statistics. Several population genetics statistics were calculated in 100-kb/10-kb and 10-kb/2-kb sliding windows and each gene in each DNA pool. Any window or gene with >50% Ns was excluded, and all statistics were based on the number of non-N nucleotides in the window. Nucleotide diversity (π, the average number of nucleotide differences per site between two DNA sequences chosen randomly from the sample population; ref. 67) was calculated using the following formula:

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{i} x_i x_j \pi_{ij}$$

Here x_i and x_j are the respective frequencies of the ith and jth sequences, π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequences, and n is the number of sequences in the sample. The Watterson estimate (θ_w; ref. 68), which is an estimation of population

mutation rate, was calculated on the basis of the number of segregating sites using the formula

$$\theta_w = \frac{S}{a_n}$$

where S is the number of segregating sites and

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima's D, calculated as described in ref. 69. F_{ST} (ref. 70) is a measure of population differentiation estimated from the average pairwise differences between chromosomes in each analysis panel compared to the combined samples as described in ref. 71

$$F_{ST} = 1 - \frac{\sum_j \binom{n_j}{2} \sum_i 2\frac{n_{ij}}{n_{ij}-1} x_{ij}(1-x_{ij}) / \sum_j \binom{n_j}{2}}{\sum_i 2\frac{n_i}{n_i-1} x_i(1-x_i)}$$

where x_{ij} is the estimated frequency of the minor allele at SNP i in population j, n_{ij} is the number of genotyped chromosomes at that position and n_j is the number of chromosomes analyzed in that population. The lack of the j subscript in the denominator indicates that statistics n_i and x_i are calculated across the combined data sets.

The relative diversity among two pooled samples was compared by a nucleotide diversity ratio (π) between the two pools for each window or gene. For example, the ratio $\pi_{MA\text{-}wild}/\pi_{MA\text{-}landrace}$ measures the relative difference in diversity between the Mesoamerican wild gene pool and the Mesoamerican landrace gene pool. Similarly, an F_{ST} value was calculated for each window and gene to compare the differentiation between any two pools.

Identifying selected windows and genes and defining sweep windows. A composite scoring system was used to determine whether a 10-kb/2-kb sliding or gene window was under selection. This approach is similar to the one applied for silk moth where a reduction in nucleotide diversity and Tajima's D was applied to discover domestication-related genes[72]. Here a 10-kb/2-kb window or a gene was considered a selection window or domestication candidate gene if it was in the upper 90% of the pool's empirical distribution for the $\pi_{wild}/\pi_{landrace}$ ratio and F_{ST} statistics. The cutoff values for various comparisons can be found in **Supplementary Table 18**. All 10-kb/2-kb selection windows within 40 kb of each other were merged in a 'sweep window'. The numbers of domestication candidates and total genes were calculated for each sweep window.

Annotating candidates for seed weight and size in common bean. We used the *Arabidopsis* protein sequence for all genes found to be associated with seed weight[43,73] as queries for a BLASTP analysis of a database of the common bean proteins. We identified 141 common bean gene models with 50% identity and 80% coverage that matched 70% of the query length, and these inherited the *Arabidopsis* names for the gene associated with seed weight.

Association mapping. In total, 271 diverse modern common bean varieties from the Mesoamerican gene pool were grown in replicated field trials by North Dakota State University, Michigan State University, the University of Nebraska and Colorado State University bean breeding programs. Each variety was genotyped with 34,799 SNPs. Missing data were imputed in fastPHASE 1.3 (ref. 74) using likelihood-based imputation. Adjusted means for seed weight data across all locations were calculated using the MIXED procedure in SAS9.3 (ref. 75), where the genotype was the fixed effect and all other factors were considered to be random.

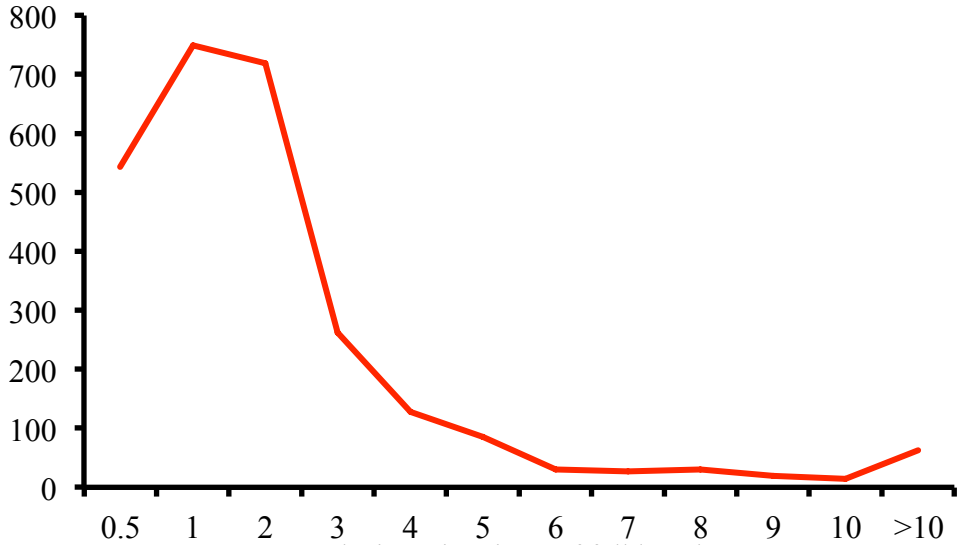
A mixed linear model (MLM) controlling for population relatedness was used to conduct the GWAS. The mixed model used was from Yu *et al.*[76], and the equation used was $y = x\beta + z\mu + \varepsilon$, where y is the seed weight phenotype, $x\beta$ indicates the genotype fixed effect, $z\mu$ represents the kinship coefficient as the random effect and ε is a vector of residual effects. An identity-by-state (IBS) kinship matrix (EMMA[77]) was used to control for population relatedness. The kinship matrix was calculated using marker loci with pairwise $r^2 > 0.5$. The linkage disequilibrium (r^2) between all marker loci was calculated in PLINK[78] using a minor allele frequency of 0.1. The EMMA kinship matrix and the GWAS were calculated in the genome association and prediction integrated tool (GAPIT) package in R[79], without P3D and compression. Only markers with minor allele frequency of 0.1 or greater were considered in the GWAS results. Protein sequences for *Arabidopsis* genes associated with seed weight[43,73] were used as queries for a BLASTP analysis against a database of common bean proteins. We identified 141 common bean gene models with 50% identity and 80% coverage that matched 70% of the query length, and these inherited the *Arabidopsis* gene names.

46. Song, Q. *et al.* Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1. 0) in soybean. *Crop Sci.* **50**, 1950–1960 (2010).
47. Van Ooijen, J. *JoinMap 4. Software for the Calculation of Genetic Linkage Maps in Experimental Populations* (Kyazma, Wageningen, The Netherlands, 2006).
48. Hyten, D.L. *et al.* High-throughput SNP discovery and assay development in common bean. *BMC Genomics* **11**, 475 (2010).
49. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
50. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
51. Schuler, G.D. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
52. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
53. Salamov, A.A. & Solovyev, V.V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
54. Yeh, R.-F., Lim, L.P. & Burge, C.B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
55. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **101**, 12404–12410 (2004).
56. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
57. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
58. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).
59. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
60. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
61. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
62. McClean, P.E. *et al.* Population structure and genetic differentiation among the USDA common bean (*Phaseolus vulgaris* L.) core collection. *Genet. Resour. Crop Evol.* **59**, 499–515 (2012).
63. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
64. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
65. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
67. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
68. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
69. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
70. Hudson, R.R., Slatkin, M. & Maddison, W. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
71. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
72. Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 433–436 (2009).
73. Kesavan, M., Song, J.T. & Seo, H.S. Seed size: a priority trait in cereal crops. *Physiol. Plant.* **147**, 113–120 (2013).
74. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
75. SAS Institute, Inc. *SAS 9.3 Language Reference: Concepts, Second Edition* (SAS Institute, Inc., Cary, NC, 2012).
76. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
77. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
78. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
79. Lipka, A.E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).

A reference genome for common bean and genome-wide analysis of dual domestications.

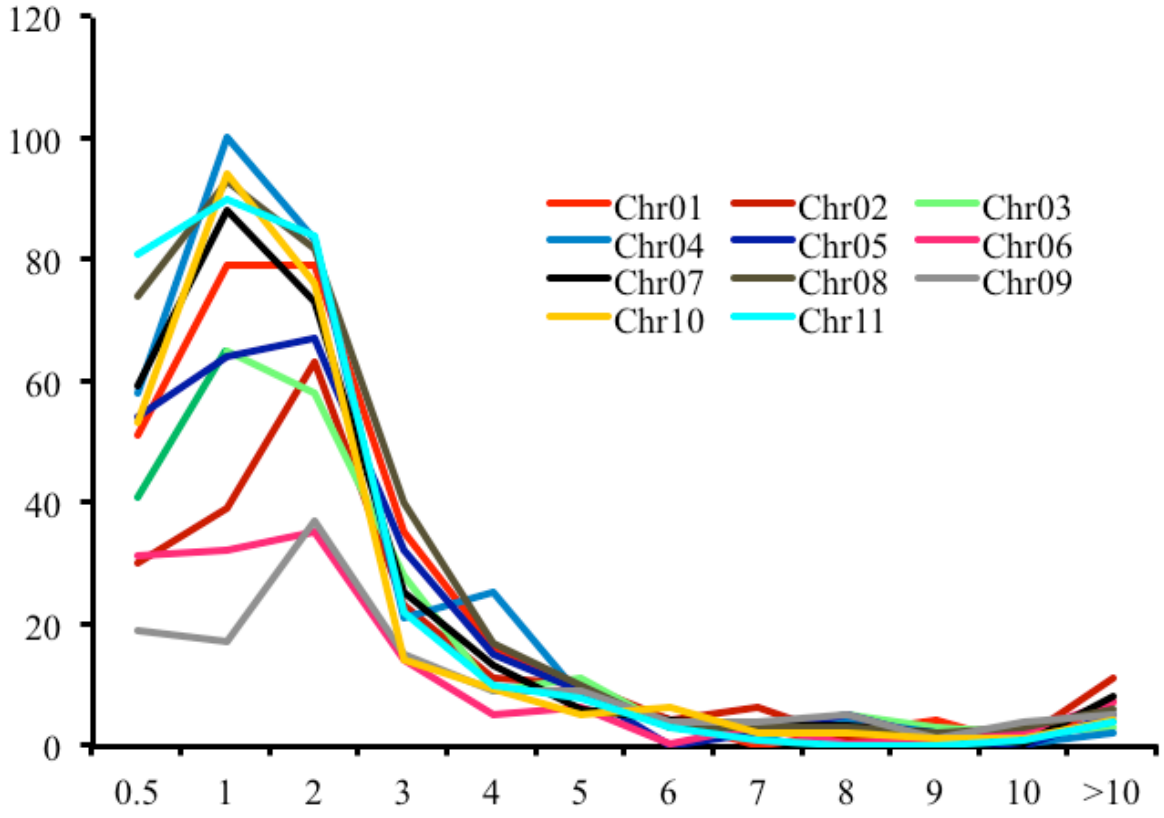
Jeremy Schmutz, Phillip McClean, Sujan Mamidi, G. Albert Wu, Steven B. Cannon, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, Qijian Song, Carolina Chavarro, Mirayda Torres-Torres, Valerie Geffroy, Samira Mafi Moghaddam, Dongying Gao, Brian Abernathy, Kerrie Barry, Matthew Blair, Mark A. Brick, Mansi Chovatia, Paul Gepts, David M Goodstein, Michael Gonzales, Uffe Hellsten, David L. Hyten, Gaofeng Jia, James D. Kelly, Dave Kudrna, Rian Lee, Manon M.S. Richard, Phillip N. Miklas, Juan M. Osorno, Josiane Rodrigues, Vincent Thareau, Carlos A. Urrea, Mei Wang, Yeisoo Yu, Ming Zhang, Rod A. Wing, Perry B. Cregan, Daniel S. Rokhsar, Scott A. Jackson

Content:



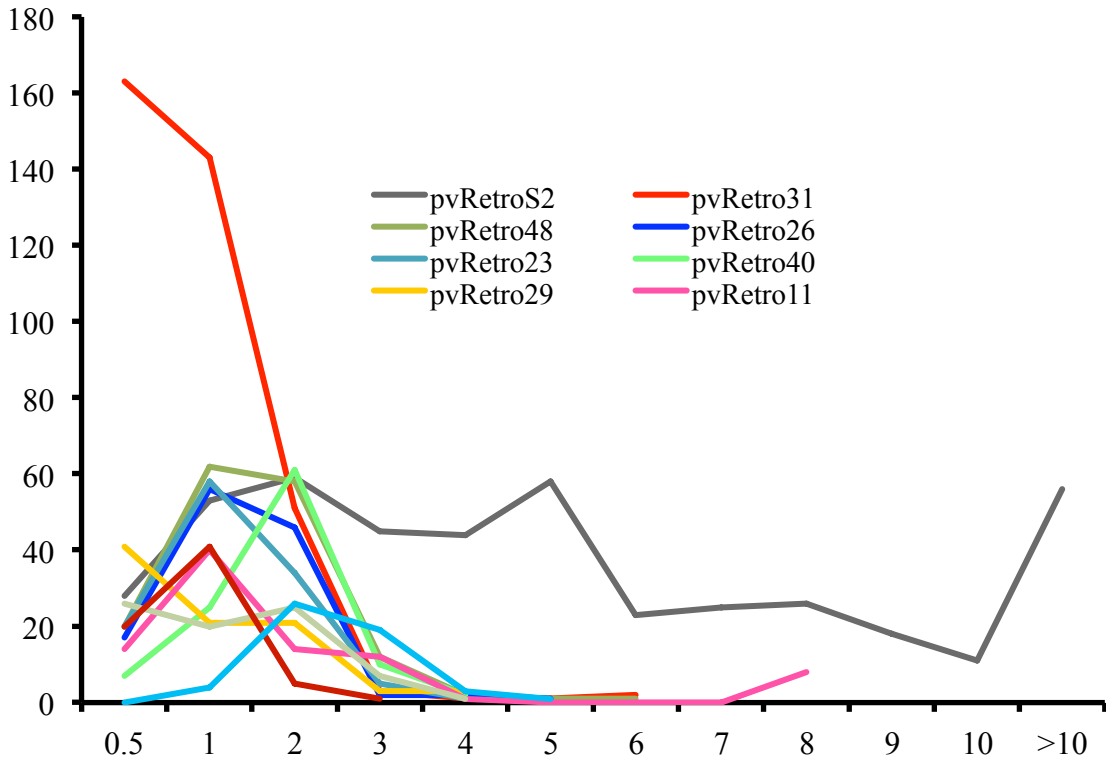
Supplementary Figure 1. The insertion times of full length LTR retrotransposons in common bean.

Nature Genetics: doi:10.1038/ng.3008

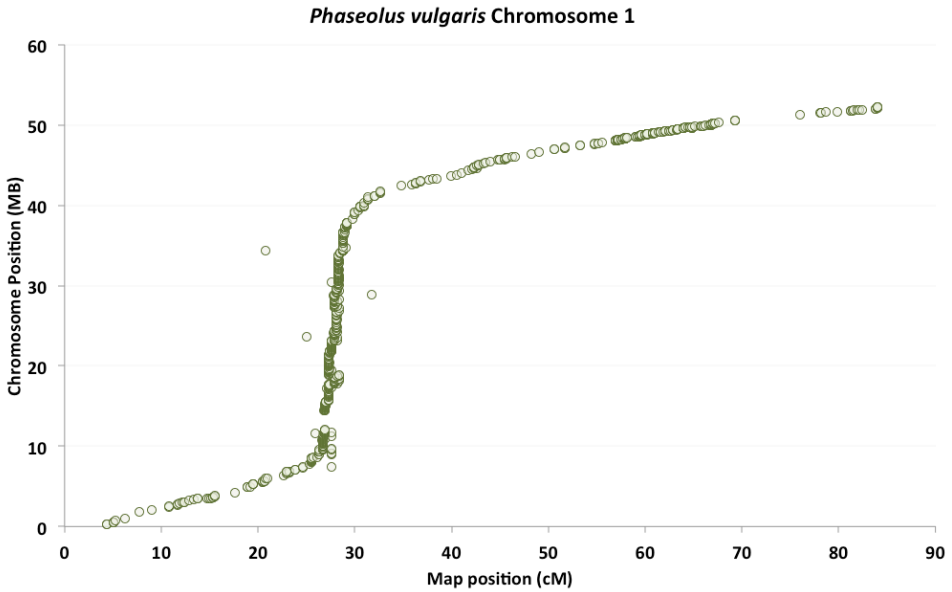


Supplementary Figure 2. The integration times of full length LTR retrotransposons on the 11 chromosomes of common bean.

Nature Genetics: doi:10.1038/ng.3008

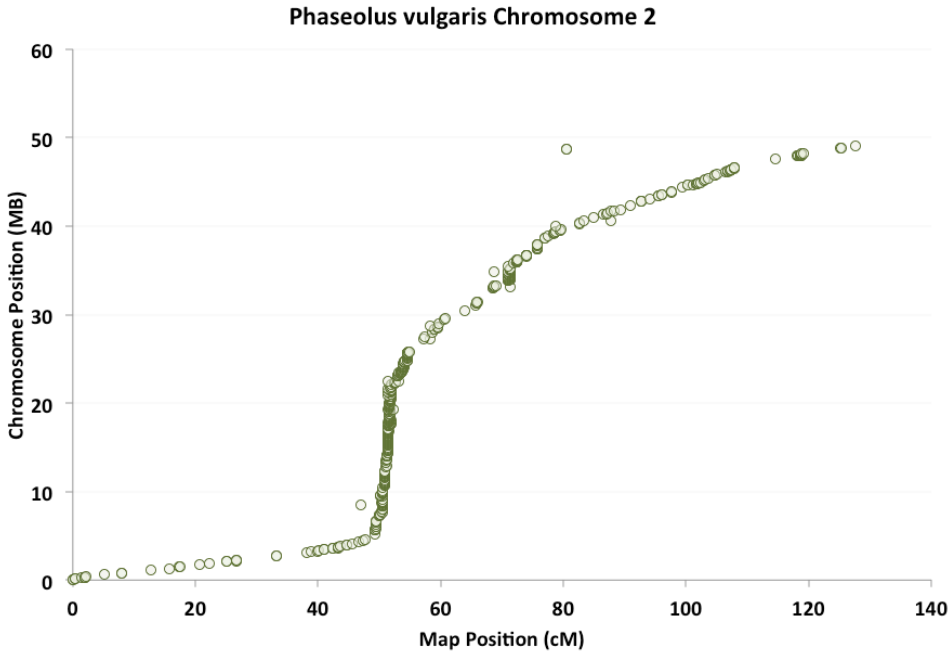


Supplementary Figure 3. The insertion times of 11 LTR retrotransposon families each of which contains more than 50 complete elements.

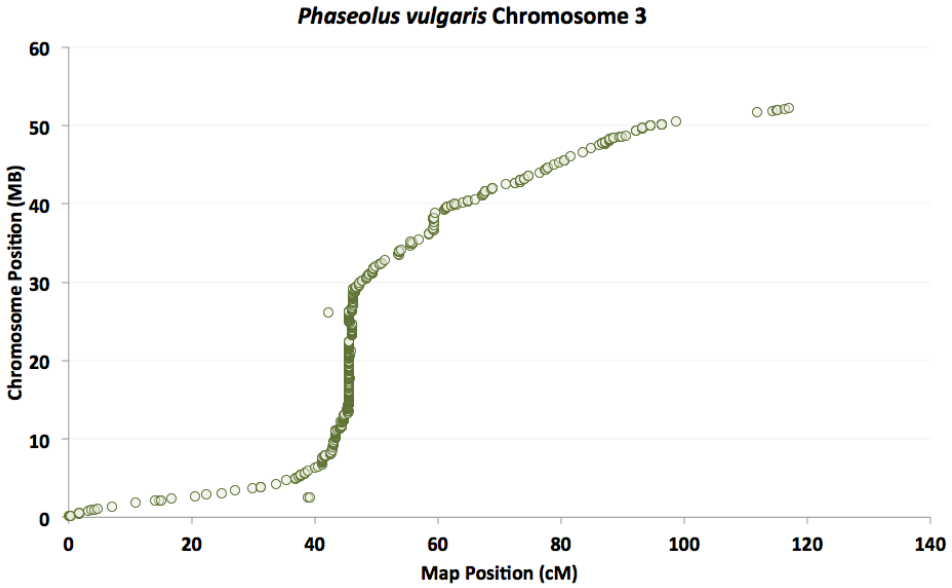


Supplementary Figure 4. Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 1.

Nature Genetics: doi:10.1038/ng.3008

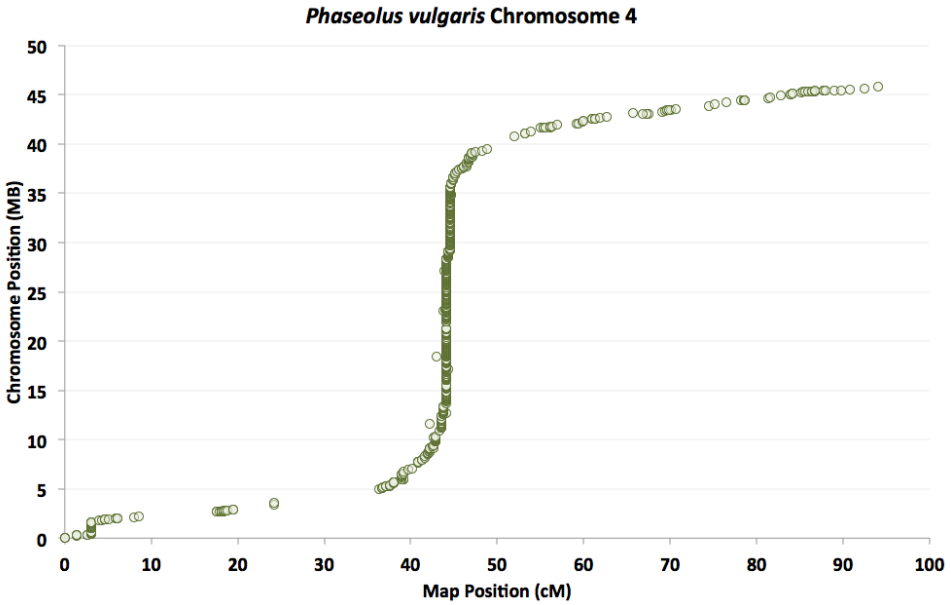


Supplementary Figure 5: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 2.

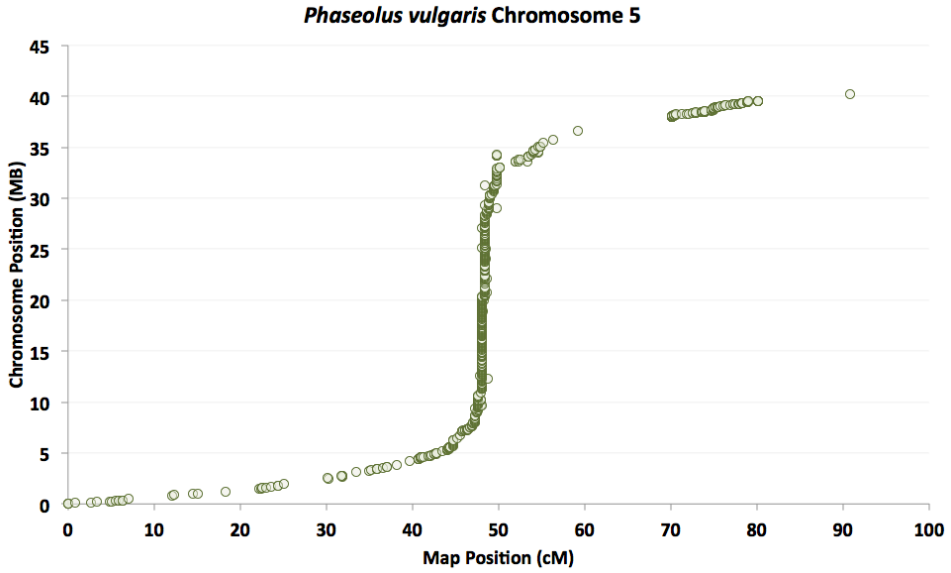


Supplementary Figure 6: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 3.

Nature Genetics: doi:10.1038/ng.3008

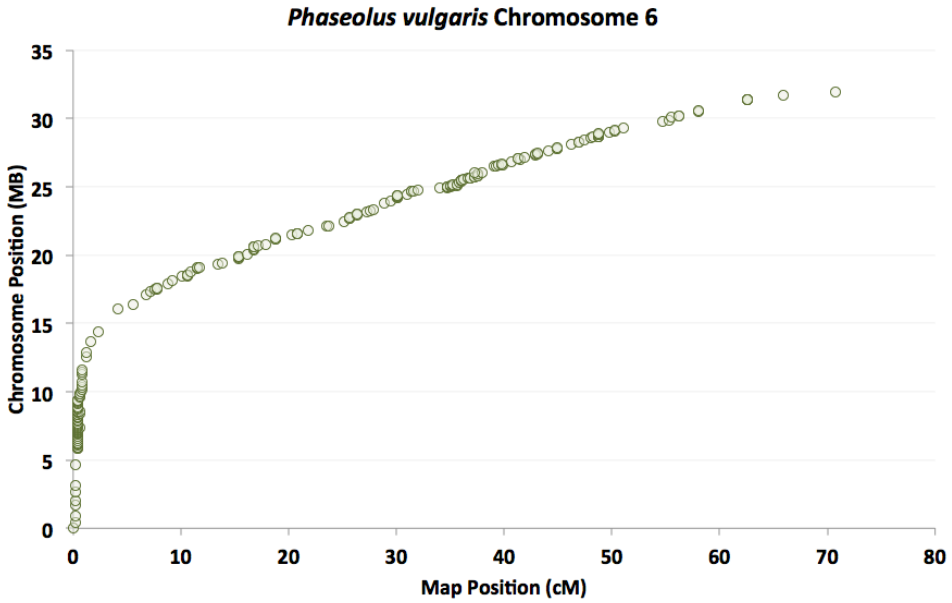


Supplementary Figure 7: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 4.

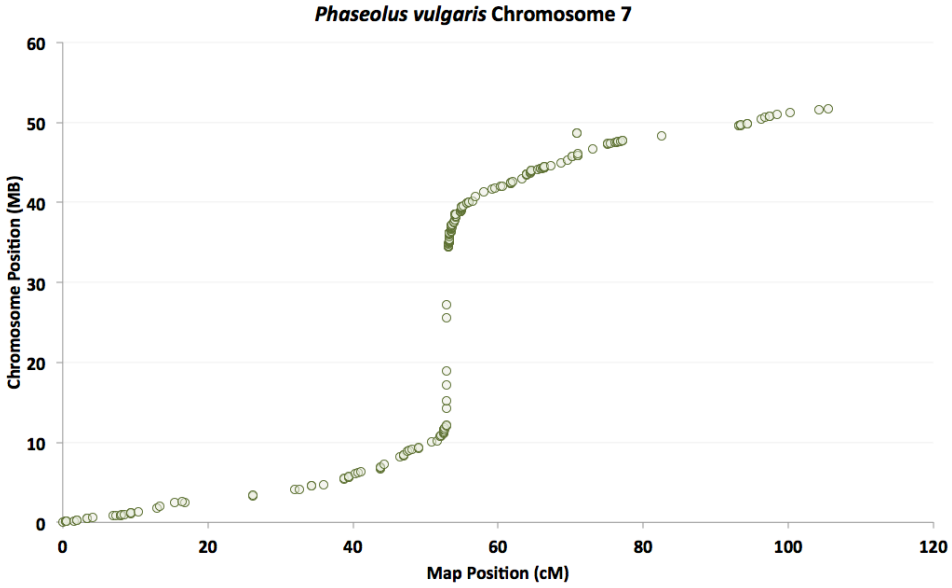


Supplementary Figure 8: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 5.

Nature Genetics: doi:10.1038/ng.3008

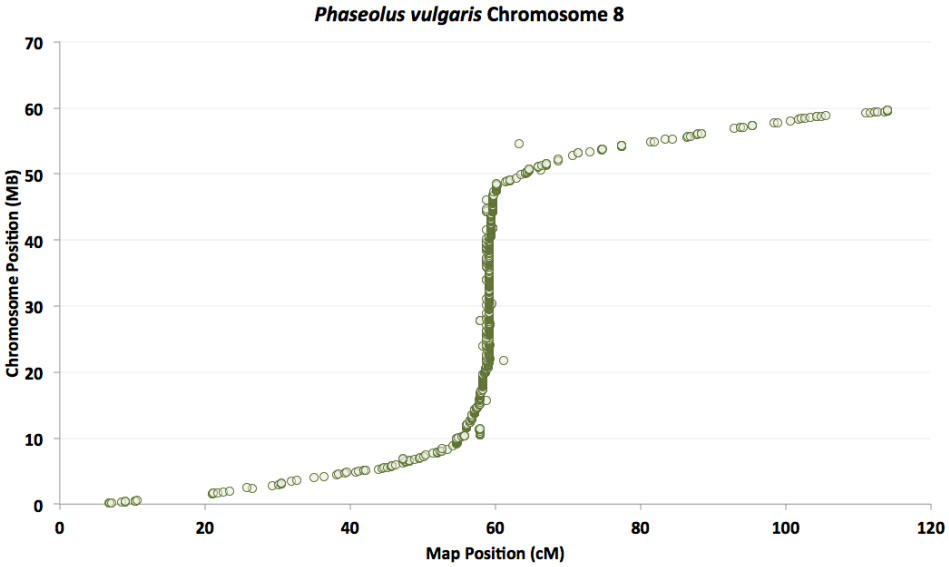


Supplementary Figure 9: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 6.

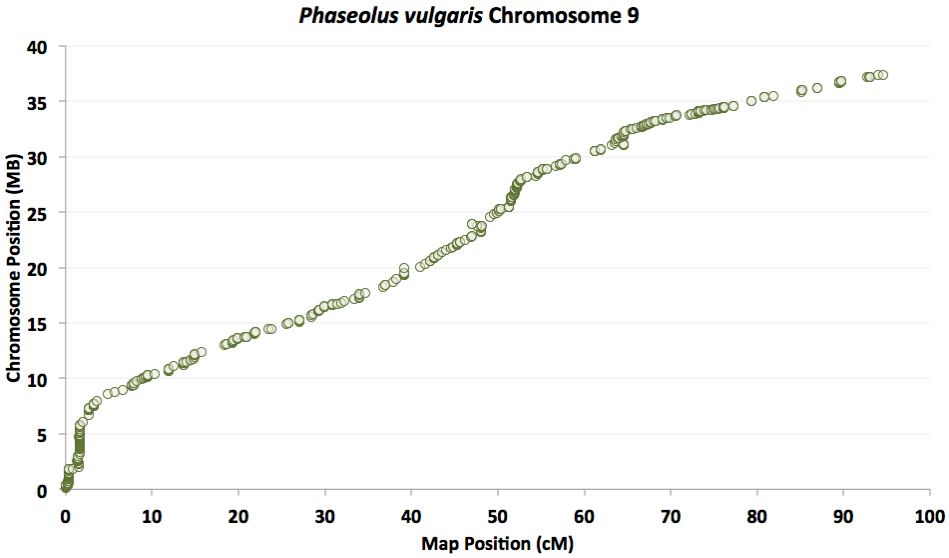


Supplementary Figure 10: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 7.

Nature Genetics: doi:10.1038/ng.3008

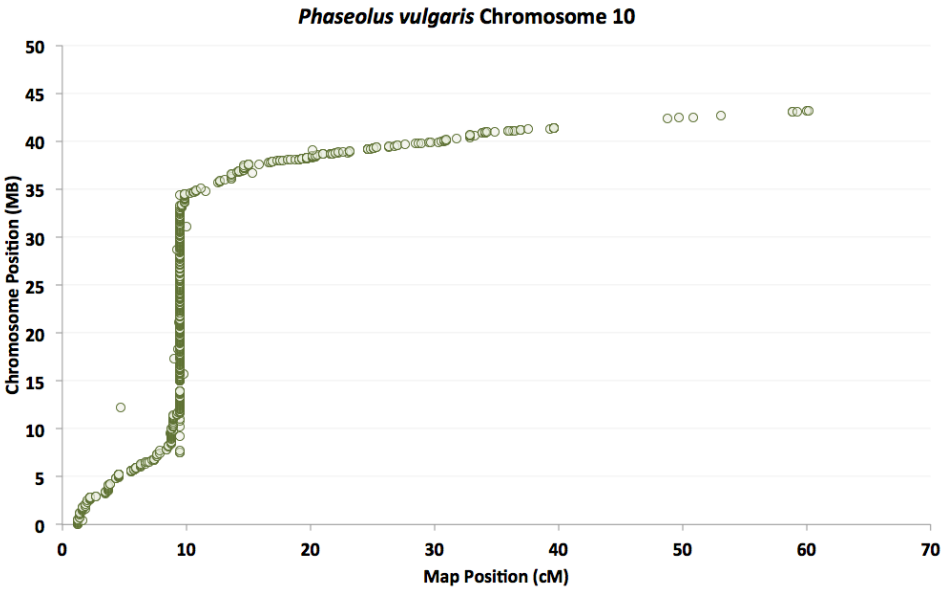


Supplementary Figure 11: Marker placements for the genetic map on the *Phaseolus vulgaris* chromosome 8.

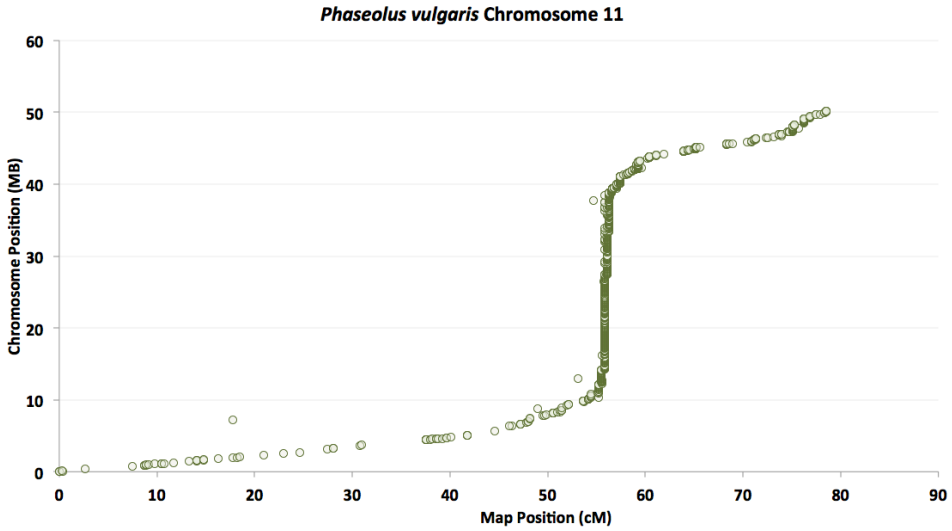


Supplementary Figure 12: Marker placements for the genetic map on *Phaseolus vulgaris* chromosome 9.

Nature Genetics: doi:10.1038/ng.3008

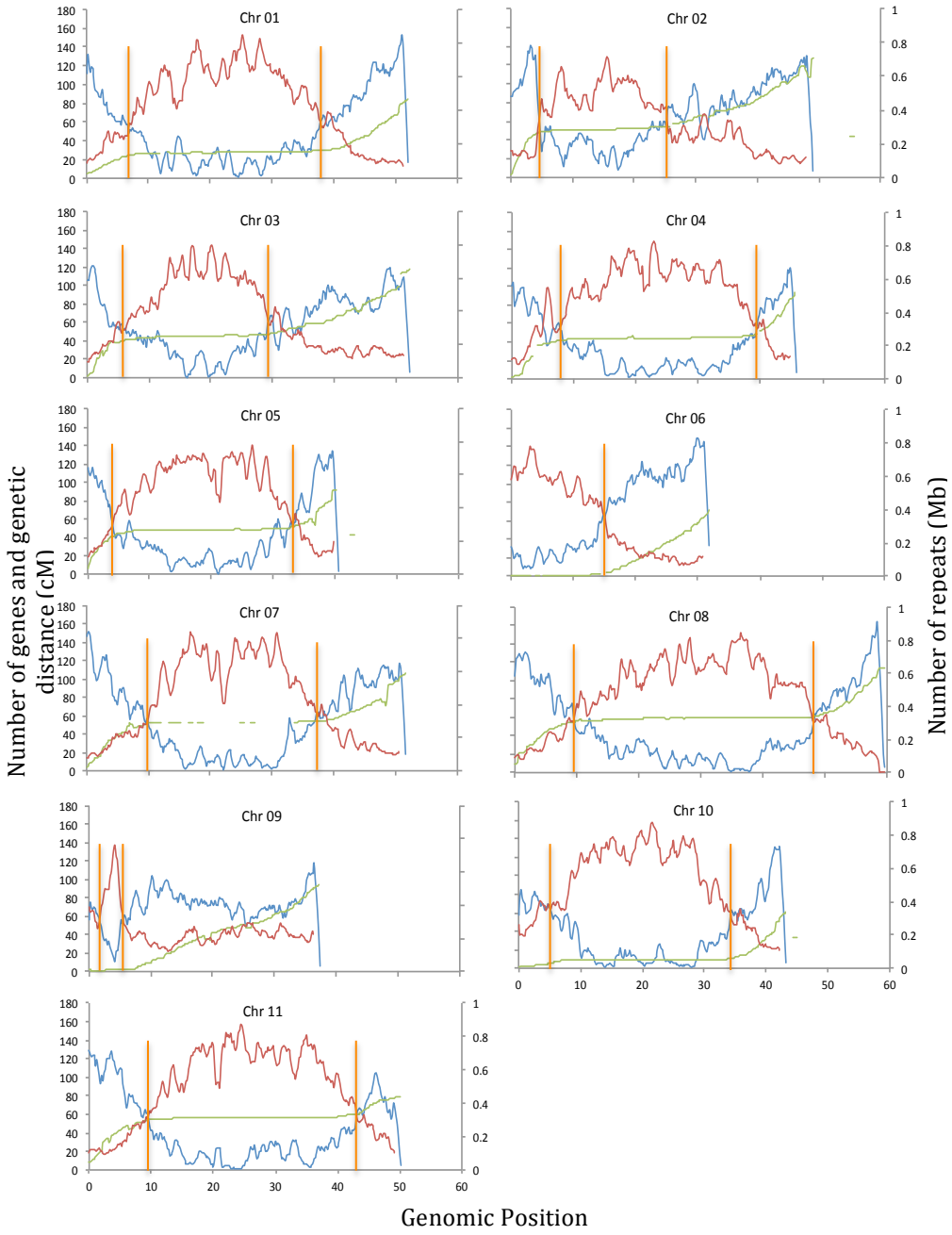


Supplementary Figure 13: Marker placements for the genetic map on *Phaseolus vulgaris* chromosome 10.



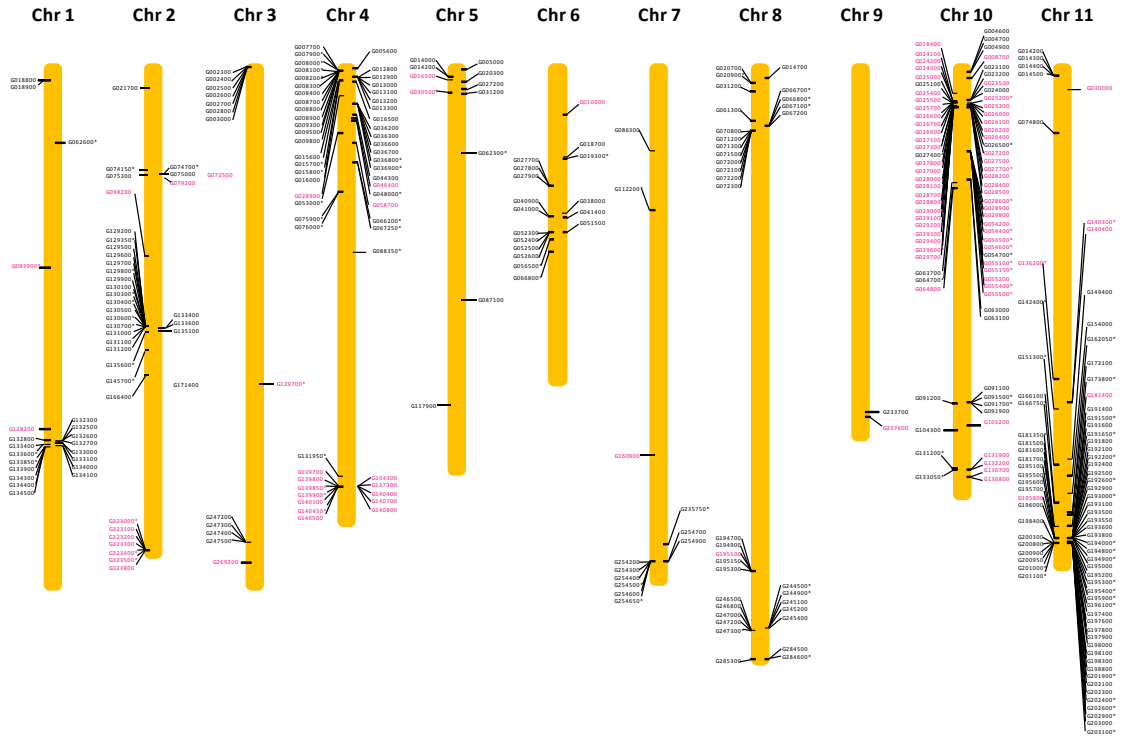
Supplementary Figure 14: Marker placements for the genetic map on *Phaseolus vulgaris* chromosome 11.

Nature Genetics: doi:10.1038/ng.3008



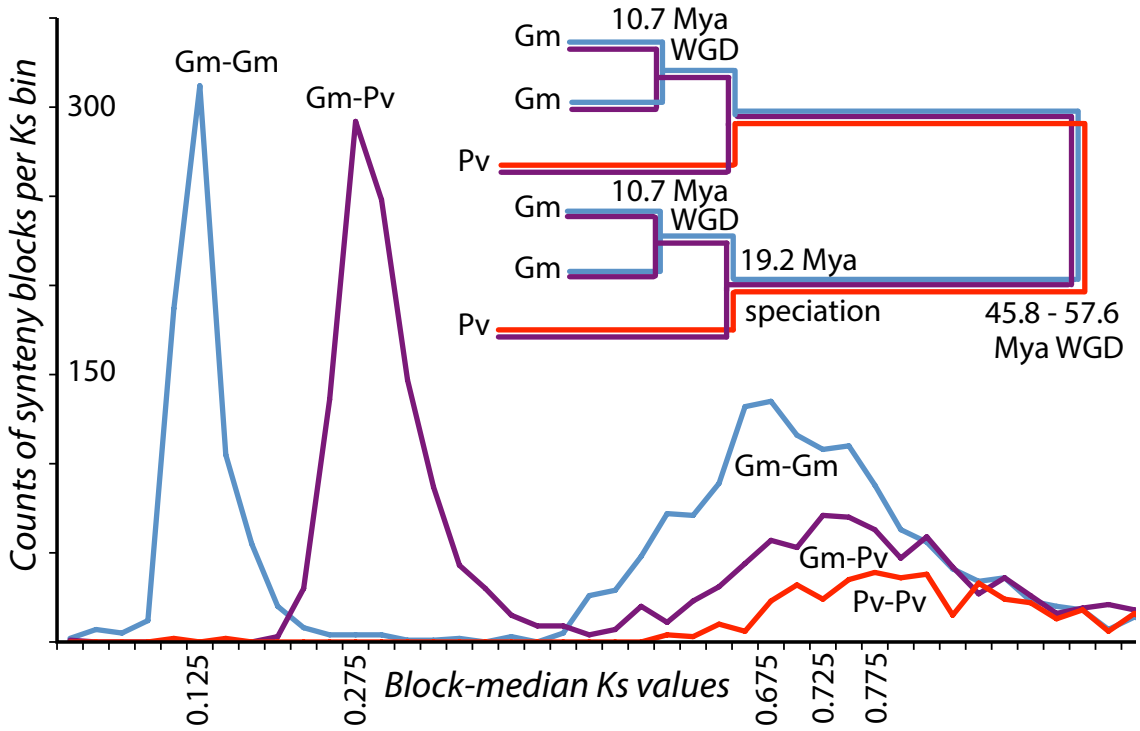
Supplementary Figure 15. Identification of pericentromeric regions. Identification of pericentromeric regions. Based on the comparison between physical distance (X axis) with gene density (blue line, left Y axis), repeats density (red line, right Y axis) and average of genetic distance (green line, left Y axis). Yellow vertical bars indicate position of transition from euchromatic arms to pericentromeres. All measures are based in a 1Mb window increasing every 200 kb. The gene density includes 27,197 genes and the genetic distance is based on 6,945 markers mapped in the Stampede x Redhawk population in a F_2 generation.

Nature Genetics: doi:10.1038/ng.3008

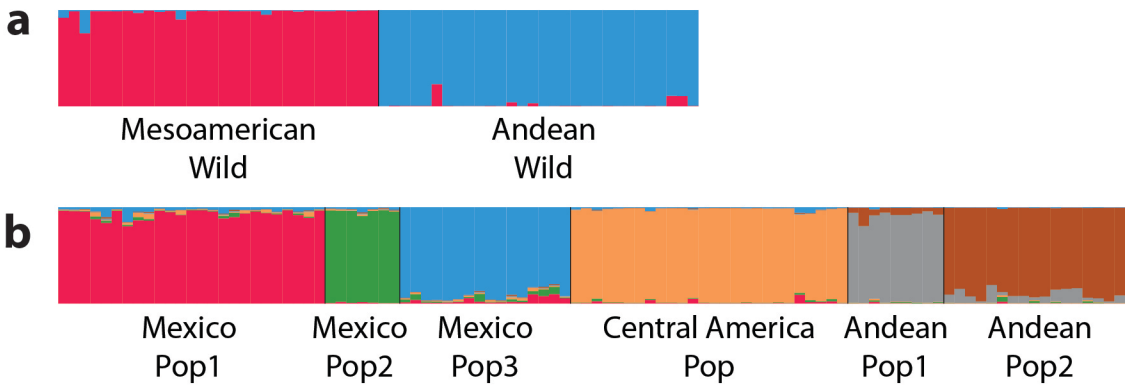


Supplementary Figure 16. Physical map of the 11 common bean chromosomes with individual CNLs and TNLs. The relative map position of 376 NL encoding genes is shown on the individual pseudomolecules depicting the chromosomes 1-11. Each gene has a unique label representing the 7 last informative digits from the annotation. For example, G002600 located on pseudomolecule 3 corresponds to the gene Phvul.003G002600. Genes encoded by the positive DNA strand are depicted on the right side of the chromosome, whereas those encoded by the negative strand are shown on the left. TNL sequences are presented in pink and CNL sequences are presented in black. NL corresponding to a pseudogene are denoted by an asterisk (*) after their name.

Nature Genetics: doi:10.1038/ng.3008

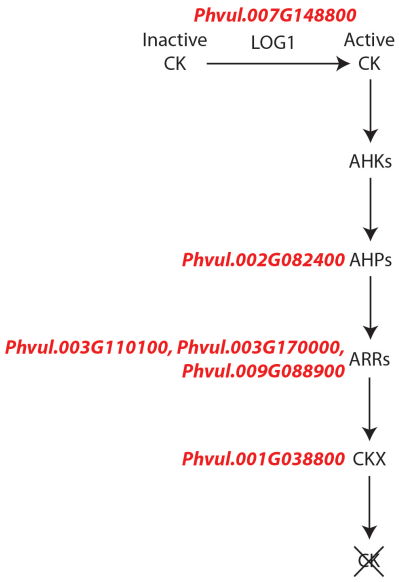


Supplementary Figure 17. Gene duplications and divergence estimates. Observed Ks values and inferred speciation and gene duplication divergence estimates, based on median synonymous substitutions values (Ks) of synteny-block-median Ks values from gene pairs from syntenic regions. A system of equations corresponding to branches on gene-pair lineages (red, blue, or purple) was used to determine the branch lengths in this gene-family model. Rates of substitutions are based on the divergence time estimated by Lavin et al. (2005) for *Phaseolus* and *Glycine* of 19.2 Mya.



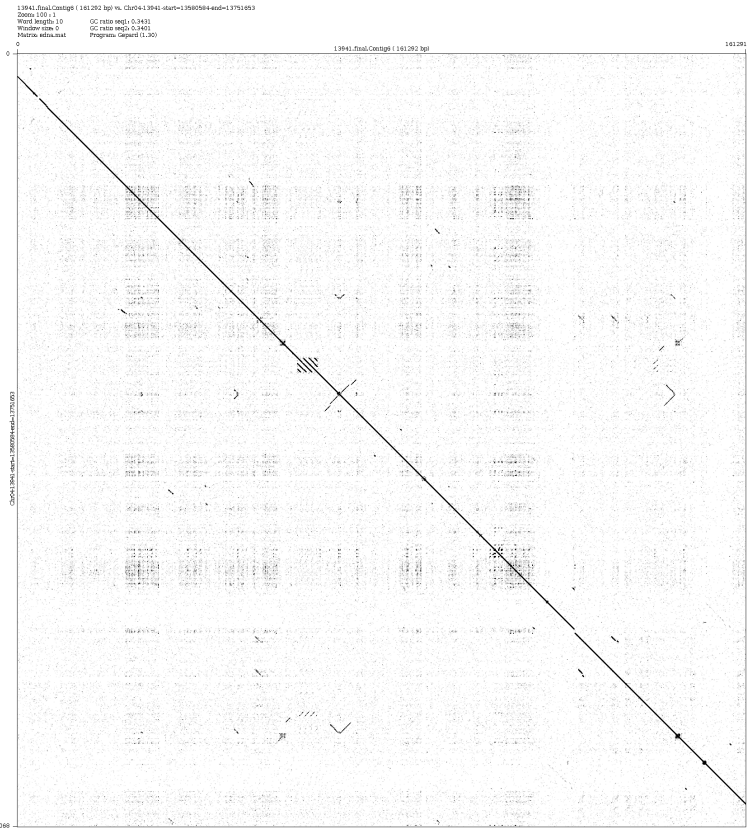
Supplementary Figure 18. Population substructure of 60 wild and 100 landrace common bean genotypes used for pooled resequencing. Population membership was defined using the STRUCTURE software. Based on historical research, the wild genotypes (a) were subdivided into two subpopulations, while the landraces (b) were defined by six subpopulations.

Nature Genetics: doi:10.1038/ng.3008

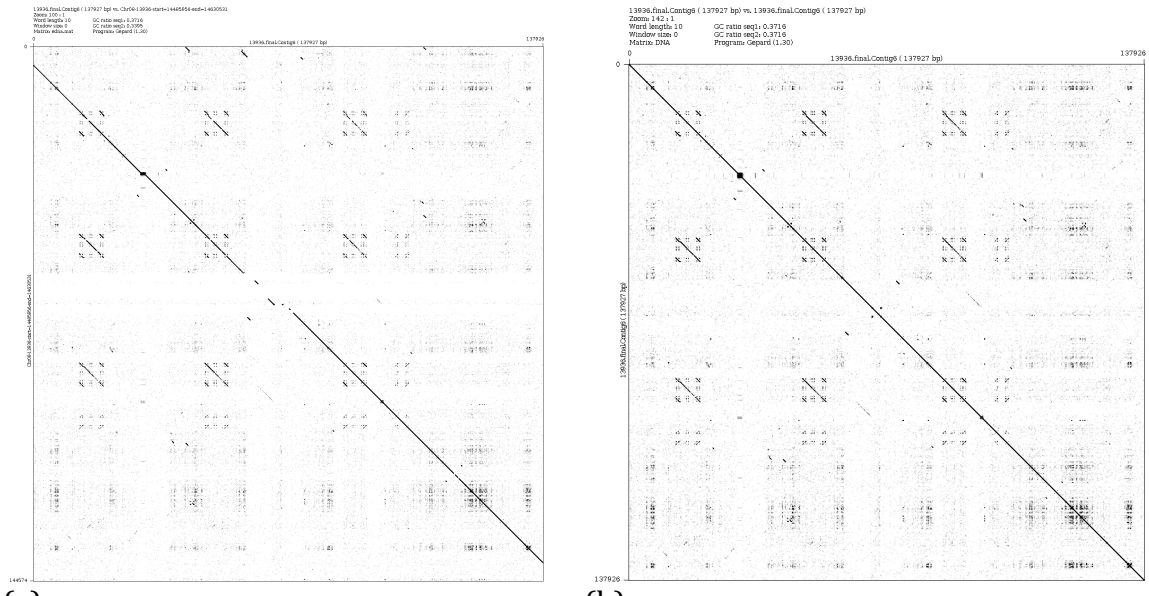


Supplementary Figure 19. Cytokinin pathway and MA domestication genes**.** Cytokinin is ynthesized from a precursor by the enzyme LOG1. It is then sensed by members of the AHK class that autophosphorylate themselves. The phosphate group is passed to AHP proteins that migrate to the nucleus and phosphorylate ARR proteins. These transcription factors in turn ctivate genes such as CKX that degrades cytokinin to modulate the effects of the hormone on ıultiple plant development processes (Hwang et al. 2012). The MA domestication candidates for genes in the pathway are noted.

Nature Genetics: doi:10.1038/ng.3008



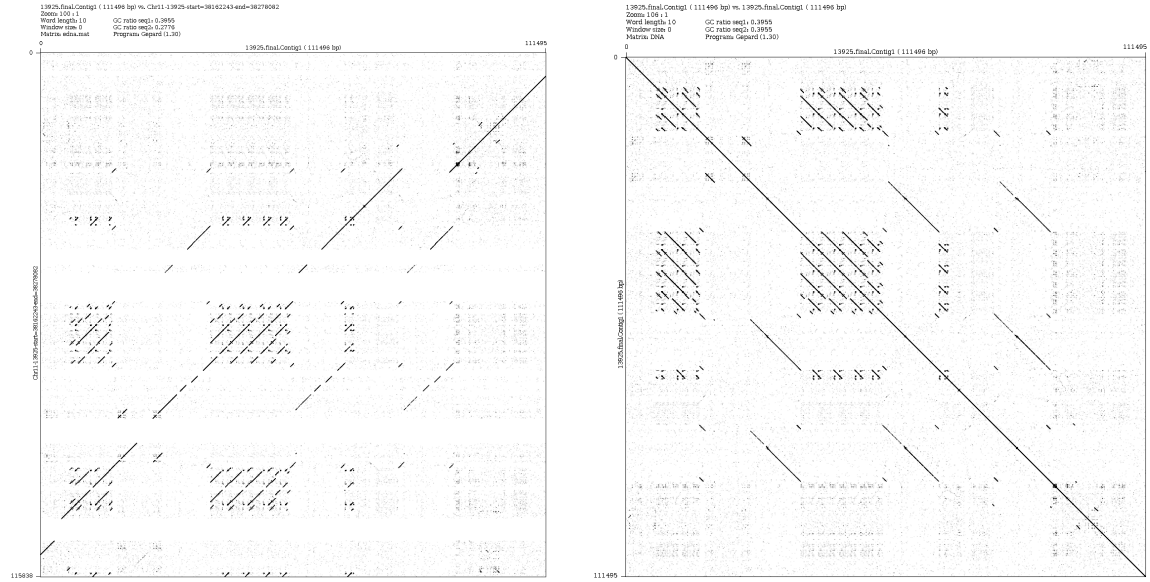
Supplementary Figure 20: Dot plot of BAC clone 13941 on a region of chromosome Pv04. This alignment is representative of the high quality BAC clone alignment.



(a) (b)

Supplementary Figure 21: Dot plot of BAC clone 13936 on a region of chromosome Pv04. This alignment is representative of a BAC clone in a moderate transposon region, where (a) is the dot plot on Chr08 and (b) is the dot plot of the clone with itself.

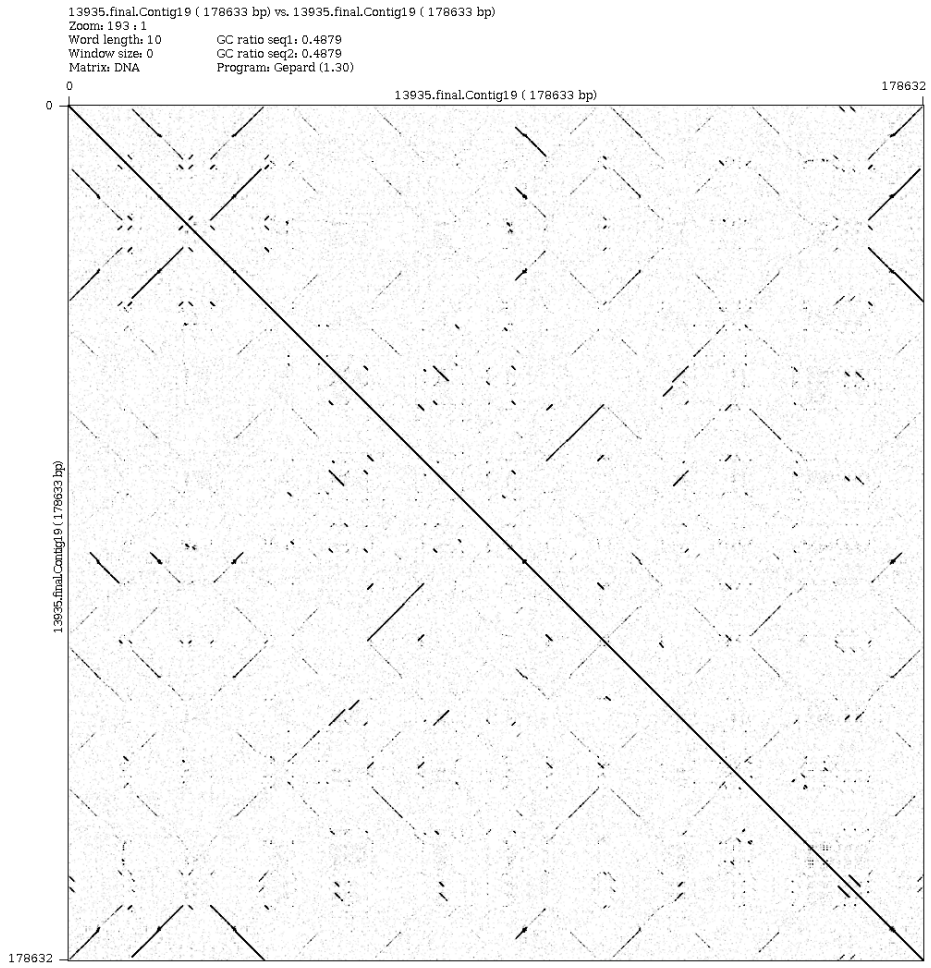
Nature Genetics: doi:10.1038/ng.3008



(a) (b)

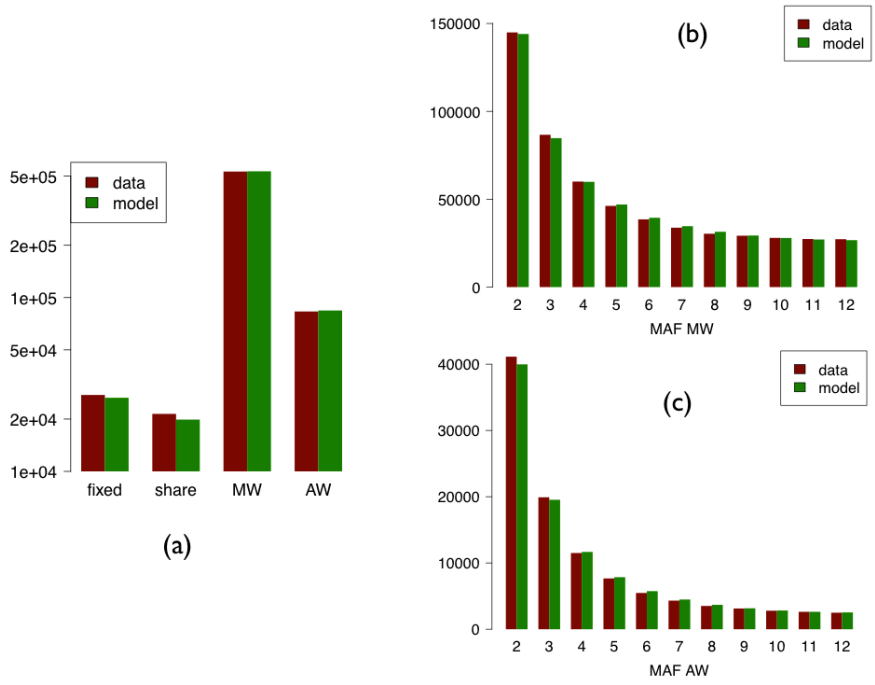
Supplementary Figure 22: Dot plot of BAC clone 13925 on a region of chromosome Pv11. This alignment is representative of a BAC clone in a dense transposon environment, where (a) is the alignment on Chr11 and (b) is the dot plot of the clone with itself.

Nature Genetics: doi:10.1038/ng.3008



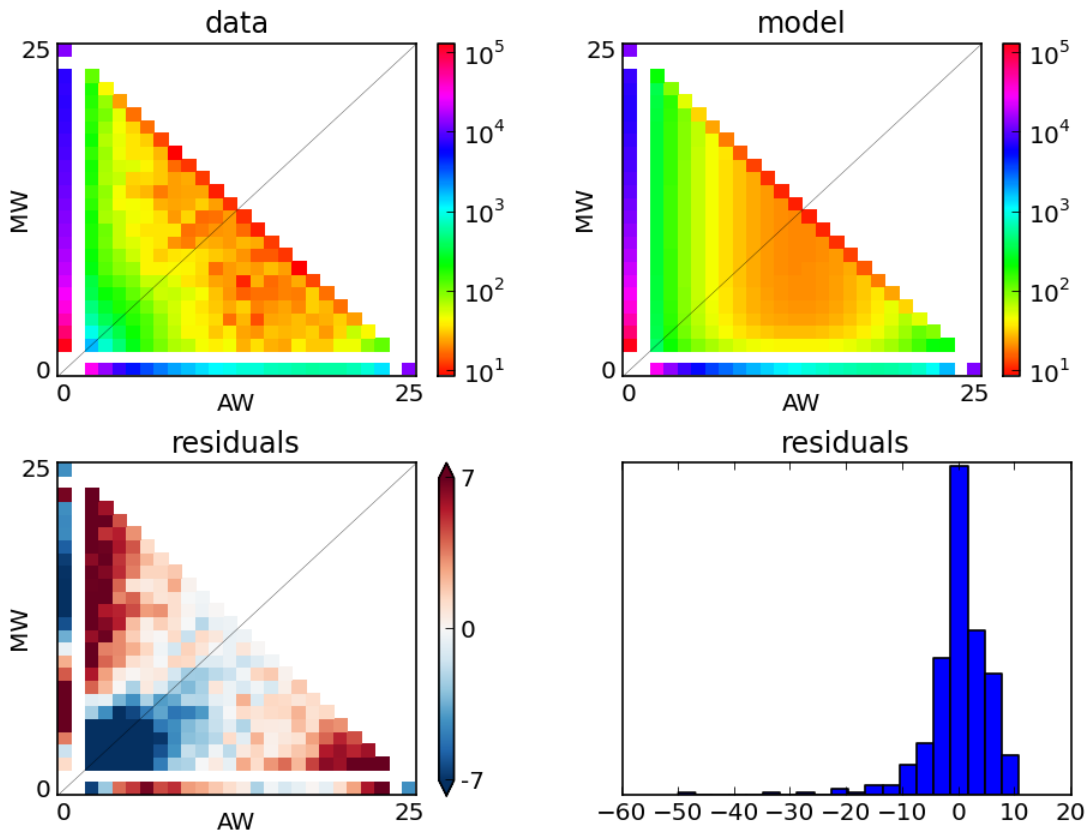
Supplementary Figure 23: Dot plot of BAC clone 13935 with itself. This clone likely resides in a region of the genome that was not resolved in the assembly.

Nature Genetics: doi:10.1038/ng.3008



Supplementary Figure 24. Data vs model prediction for the genetic variation in the two wild bean pools of wild Mesoamerican (MW) and wild Andean (AW). Each pool has been down-sampled to 25 chromosomes. Singletons in both pools have been excluded in model inference and prediction. (a) The summary statistic of 4 types of mutually exclusive single nucleotide variants: fixed=sites with 2 alleles separately fixed in MW and AW, share=sites variant in both pools, MW=sites variant in MW only, AW=sites variant in AW only. (b) Minor allele frequency distribution for the MW pool. (c) Minor allele frequency distribution in the AW pool with sample size 25.

Nature Genetics: doi:10.1038/ng.3008



Supplementary Figure 25. Joint allele frequency spectrum for the two wild pools of common bean. The pooled data were down-sampled to 25 chromosomes for each pool, and singletons were excluded both in model inference and prediction. The Anscombe residuals between the best fit model and data are shown in the bottom row. See test for discussions. MW=wild Mesoamerican, AW=wild Andean.

Nature Genetics: doi:10.1038/ng.3008

B. Supplementary Tables

Library	Sequencing Platform	Average Read Length/Insert Size	Read Number	Assembled Sequence Coverage
Linear	454 XLR & FLX+	362*	38,107,155	18.64x
GPNB	454 XLR paired	2,798 ± 1,047	589,346	0.11x
GGAS	454 XLR paired	3,922 ± 643	1,940,576	0.41x
GXSF	454 XLR paired	3,991 ± 337	467,414	0.07x
HYFA	454 XLR paired	4,729 ± 497	1,648,022	0.25x
HYFC	454 XLR paired	4,736 ± 504	1,491,648	0.24x
HYFB	454 XLR paired	4,759 ± 528	1,196,104	0.17x
HXTI	454 XLR paired	8,022 ± 1,016	1,364,808	0.22x
GXNX	454 XLR paired	9,192 ± 1,058	878,832	0.16x
HXWF	454 XLR paired	11,903 ± 1,928	724,196	0.13x
HXWH	454 XLR paired	12,231 ± 1,902	413,396	0.08x
VUK	Sanger	34,956 ± 4,536	240,384	0.20x
VUL	Sanger	36,001 ± 4,632	88,320	0.08x
PVC	Sanger	121,960 ± 16,572	81,408	0.08x
PVA	Sanger	126,959 ± 25,658	89,017	0.09x
PVB	Sanger	135,292 ± 21,487	92,160	0.09x
Total			49,412,786	21.02x

Supplementary Table 1. Genomic libraries included in the *Phaseolus vulgaris* genome assembly and their respective assembled sequence coverage levels in the final release. *Indicates that the number reported in the table is the average read length, not insert size.

Nature Genetics: doi:10.1038/ng.3008

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Scaffold Size	Basepairs	% Non-gap Basepairs
5 Mb	36	24,903	310,700,332	284,755,606	91.65%
2.5 Mb	65	33,373	418,546,348	382,376,028	91.36%
1 Mb	109	38,683	497,761,392	454,793,715	91.37%
500 Kb	122	39,252	507,057,583	463,387,171	91.39%
250 Kb	136	39,730	512,032,524	466,907,449	91.19%
100 Kb	157	40,017	515,361,076	468,917,527	90.99%
50 Kb	171	40,169	516,398,703	469,738,390	90.96%
25 Kb	213	40,452	517,980,937	470,824,917	90.90%
10 Kb	289	40,740	519,103,479	471,766,339	90.88%
5 Kb	479	41,194	520,388,386	472,773,109	90.85%
2.5 Kb	641	41,453	521,017,136	473,245,231	90.83%
1 Kb	1,100	41,920	521,675,054	473,897,487	90.84%
0 bp	1,627	42,447	522,065,413	474,287,846	90.85%

Supplementary Table 2. Summary statistics of the output of the whole genome shotgun assembly prior to screening, removal of organelles and contaminating scaffolds and chromosome-scale pseudomolecule construction. The table shows total contigs and total assembled basepairs for each set of scaffolds greater than the size listed in the left hand column.

Nature Genetics: doi:10.1038/ng.3008

Scaffold total	708
Contig total	41,391
Scaffold sequence total	521.1 Mb
Contig sequence total	472.5 Mb (1.7% gap)
Scaffold N50/L50	5/50.4 Mb
Contig N50/L50	3,273/39.5 Kb

Supplementary Table 3. Final summary assembly statistics for chromosome scale assembly.

Nature Genetics: doi:10.1038/ng.3008

Resource type	Tissue Type	Number of reads	GSNAP (Wu and Nacu 2010) Aligned	Percent Aligned
Sanger	Mixed	79,630	-	-
Illumina 2x100 bp	Roots 10 DAP (days after planting)	65,429,570	59,846,373	92.1%
Illumina 2x100 bp	Roots 19 DAP	46,593,274	44,116,235	94.9%
Illumina 2x100 bp	Nodules 19 DAP	71,716,844	66,112,750	92.7%
Illumina 2x100 bp	Stem 10 DAP	40,933,844	38,196,918	93.6%
Illumina 2x100 bp	Stem 19 DAP	61,842,390	44,116,235	94.9%
Illumina 2x100 bp	Primary leaves 10 DAP	68,255,918	61,371,430	90.5%
Illumina 2x100 bp	Young trifoliates 19 DAP	66,127,642	60,209,317	91.6%
Illumina 2x100 bp	Flower buds	68,363,986	61,332,231	90.5%
Illumina 2x100 bp	Whole Flowers	66,112,818	62,126,340	94.7%
Illumina 2x100 bp	Young pods 1-5cm seedless	66,133,582	62,301,836	94.8%
Illumina 2x150 bp	Green mature pods 11.5-13.5 cm	120,724,870	113,736,673	94.6%
Total RNA-Seq		742,234,738	687,643,736	93.2%

Supplementary Table 4. Transcript resources used for annotation for *Phaseolus vulgaris*.

Nature Genetics: doi:10.1038/ng.3008

Primary loci	27,197
Alternative transcripts	4,491
Average number of exons	5.5
Median exon length	160
Median intron length	200
Complete genes	26,279
Incomplete genes with start codon	225
Incomplete genomes with stop codon	657

Supplementary Table 5. Annotation results

Nature Genetics: doi:10.1038/ng.3008

	Centromeric regions				Pericentromeric regions		
Chr	**Start**	**End**	**Mid point**	**Range**	**Start**	**End**	**Range**
Pv1	12.2	19.9	16.1	7.7	6.8	38.0	31.2
Pv2	5.4	10.0	7.7	4.6	4.5	25.5	21.0
Pv3	14.8	16.9	15.8	2.1	6.0	29.5	23.5
Pv4	15.7	22.2	19.0	6.5	8.0	39.5	31.5
Pv5	15.3	22.7	19.0	7.5	4.0	33.8	29.8
Pv6	2.6	2.7	2.7	0.1	0.0	15.0	15.0
Pv7	16.7	30.3	23.5	13.6	9.8	37.5	27.7
Pv8	24.3	38.2	31.2	13.9	9.8	48.0	38.2
Pv9	1.5	5.8	3.7	4.3	1.5	5.8	4.3
Pv10	30.6	31.3	31.0	0.7	5.2	34.0	28.8
Pv11	16.0	17.1	16.6	1.0	9.8	43.0	33.2

Supplementary Table 6. Start and end points of centromeric regions in Mb based on BLASTN of CentPv1 and CentPv2 repeats. Start and end point of pericentromeric regions in Mb identified following the plots on Supplementary Figure 15.

Nature Genetics: doi:10.1038/ng.3008

	Genetic Length (cM)	Physical length (Kb)	Kb/cM Pericentromere	Kb/cM euchromartic arms	Kb/cM per chromosome
Chr01	84.0	52183.5	5210	278	651.5
Chr02	127.6	49033.7	3084	233	384.2
Chr03	116.9	52218.6	3452	262	445.6
Chr04	94.0	45793.2	4701	164	486.7
Chr05	90.8	40237.5	2342	134	443.0
Chr06	70.8	31973.2	6102	239	451.3
Chr07	105.4	51698.4	9179	233	489.7
Chr08	114.0	59634.6	6913	208	554.7
Chr09	94.6	37399.6	3322	352	394.0
Chr10	60.2	43213.2	5388	267	732.1
Chr11	78.5	50203.6	5877	232	638.9
Mean	**94.3**	**46689.9**	**5052**	**237**	**515.6**

Supplementary Table 7. Physical (Kb) and genetic (cM) position of the last marker mapped in each chromosome and recombination rate (Kb/cM) per chromosome and comparison between pericentromeric regions and euchromatic arms.

Nature Genetics: doi:10.1038/ng.3008

Super families of TEs	Number of TEs ($X10^3$)	Coverage of TEs (bp)	Fraction of genome (%)
Class 1	281.3	185,960,175	39.36
LTR retrotransposon	242.9	173,201,891	36.66
Ty3-gypsy	145.1	118,698,650	25.12
Ty1-copia	61.2	44,242,298	9.37
others	36.6	10,260,943	2.18
LINEs	37.5	12,599,869	2.67
SINEs	1.0	158,415	0.03
Class 2	87.1	25,979,571	5.50
CACTA	43.9	12,726,168	2.69
Harbinger/PIF	0.5	264,755	0.06
hAT	3.9	1,028,733	0.22
Helitron	18.2	5,037,722	1.07
MULE	20.6	6,922,193	1.46
Unclassified TEs	14.7	2,680,413	0.57
Total	383.2	21,4620,159	45.42

Supplementary Table 8. Summary of transposable elements (TEs) in *Phaseolus vulgaris*.

Nature Genetics: doi:10.1038/ng.3008

Chromosomes	Sizes (bp)	Retrotransposons			DNA transposons			All transposons		
		Number ($X10^3$)	Coverage (bp)	Fraction (%)	Number ($X10^3$)	Coverage (bp)	Fraction (%)	Number ($X10^3$)	Coverage (bp)	Fraction (%)
Chr1	47,951,512	28.8	18,720,056	39.0	8.5	2,566,290	5.35	38.8	21,562,140	45.0
Chr2	45,612,791	23.4	12,946,998	28.39	8.2	2,299,629	5.04	33.2	15522579	34.03
Chr3	48,183,191	25.5	15,542,892	32.26	9.3	2,614,464	5.43	36.6	18,485,093	38.36
Chr4	41,235,254	27.6	18,829,670	45.67	8.3	2,518,111	6.11	37.0	21,546,355	52.25
Chr5	36,832,944	24.2	16,333,307	44.24	6.9	2,106,756	5.72	32.3	18,644,314	50.62
Chr6	29,540,412	15.6	9,103,879	30.82	5.3	1,524,163	5.16	21.8	10,796,910	36.55
Chr7	47,238,532	26.5	18,256,201	38.65	7.9	2,347,605	4.97	35.82	20,861,506	44.16
Chr8	53,869,949	34.2	23,124,655	42.93	9.6	2,880,029	5.35	45.3	26,275,208	48.78
Chr9	35,032,714	16.3	7,273,948	20.76	7.2	1,827,118	5.22	25.2	9,372,017	26.75
Chr10	38,134,409	25.7	18,585,806	48.74	7.1	2,254,243	5.91	33.7	21,012,489	55.10
Chr11	44,565,803	29.8	20,779,963	36.63	8.0	2,466,887	5.54	39.1	23,489,119	52.71

Supplementary Table 9. Transposon distribution across the 11 chromosomes of *Phaseolus vulgaris*.

Nature Genetics: doi:10.1038/ng.3008

	#Full length	#Pseudo	#Total	%
TNL	**82**	**24**	**106**	**28.2%**
TIR-NB-LRR	73	20	93	
TIR-NB	9	4	13	
CNL	**185**	**85**	**270**	**71.8%**
CN	3	1	4	
N	5	2	7	
NL	91	64	155	
CNL	86	18	104	
#Total	267	109	**376**	

Supplementary Table 10. Numbers of common bean genes that encodes domains similar to plant R proteins

Nature Genetics: doi:10.1038/ng.3008

	Pv total genes	Pv synteny genes	Gm A	Gm B	Lost in A or B	Retained in A or B	% Gm A	%Gm B	Ratio A:B
Pv01	2694	2116	1971	1888	373	1743	93%	89%	1.04
Pv02	3338	2695	2426	2451	513	2182	90%	91%	0.99
Pv03	2973	2294	2112	1894	582	1712	92%	83%	1.12
Pv04	1789	1035	908	902	260	775	88%	87%	1.01
Pv05	1863	1349	1198	1139	361	988	89%	84%	1.05
Pv06	2221	1649	1508	1417	373	1276	91%	86%	1.06
Pv07	2812	2146	1961	1920	411	1735	91%	89%	1.02
Pv08	2932	2067	1873	1810	451	1616	91%	88%	1.03
Pv09	2633	2134	1947	1945	376	1758	91%	91%	1.00
Pv10	1659	1020	933	890	217	803	91%	87%	1.05
Pv11	2168	1274	1177	1055	316	958	92%	83%	1.12
Total	27082	19779	18014	17311	4233	15548			
Mean	2462	1798	1638	1574	385	1413			

Supplementary Table 11. *Phaseolus* synteny genes and their corresponding chromosomes in *Glycine*. Lost and retained genes in *Glycine* homolog chromosomes with based on *Phaseolus* genes. (% of GmA and GmB were calculated over the total number of genes in synteny blocks per chromosome.)

Nature Genetics: doi:10.1038/ng.3008

Pool definition (abbreviation)	Genepool	Pool size	Sequence collected in GB	Diploid genome equivalents
Landrace Mexico 1	Meso	25	153.2	6.1x
Landrace Mexico 2	Meso	7	47.9	6.8x
Landrace Mexico 3	Meso	16	102.9	6.4x
Landrace Central America	Meso	26	136.8	5.3x
Landrace South Andes	Adean	9	63.7	7.1x
Landrace North Andes	Andean	17	57.2	3.4x
Wild Mesoamerican	Meso	30	161.5	5.4x
Wild Andean	Andean	30	147.4	4.9x

Supplementary Table 12. *Phaseolus vulgaris* race and wild pool resequencing

Nature Genetics: doi:10.1038/ng.3008

	95% confidence intervals	un-bootstrapped fit
$M12=2*N_{anc}*m12$	0.072 - 0.1	0.087
$M21=2*N_{anc}*m21$	0.12 - 0.152	0.135
Ancestral population size	158900 - 176200	1.68E+05
Divergence time (yr)	146200 - 183700	1.65E+05
MW initial pop size	124900 - 205800	1.55E+05
MW final pop size	463300 - 658300	5.61E+05
AW bottleneck pop size	2304 - 8978	3.59E+03
AW bottleneck duration (yr)	60370 - 99470	7.59E+04
AW final effective pop size	188500 - 271300	2.19E+05
AW exponential growth duration	65500 - 99150	8.88E+04

Supplementary Table 13. Demographic model parameters for the divergence of the wild Mesoamerican and wild Andean bean pools. The confidence intervals were derived from 100 bootstrap replicates. A population size refers to the effective population size. For example, MW initial population size refers to the effective population size of the wild Mesoamerican pool right after its split from the wild Andean (AW) pool. M12 is the AW to MW population migration rate, and M21 is the MW to AW migration rate. A base substitution rate of 8.46e-9 /bp/yr is used. See Fig. 1 for model illustration and text for details.

Nature Genetics: doi:10.1038/ng.3008

Population	Population size	# of SNPs in population	# of SNPs in genes	% of SNPs in genes
Middle American				
Wild	30	8,890,318	1,422,926	16.01
All landraces	74	9,661,807	1,487,930	15.40
Mexican landraces	48	9,420,133	1,460,670	15.51
Mexican sub population 1	25	6,065,384	949,620	15.66
Mexican sub population 2	7	5,843,761	971,569	16.63
Mexican sub population 3	16	7,009,370	1,113,682	15.89
Central America sub population	26	5,046,476	808,411	16.02
Andean				
Wild	30	2,837,493	422,393	14.89
All landraces	26	3,154,648	522,897	16.58
Andean sub population 1	9	1,397,405	221,196	15.83
Andean sub population 2	17	2,589,280	439,086	16.96

Supplementary Table 14. SNP diversity among pooled sequencing populations.

Nature Genetics: doi:10.1038/ng.3008

	100kb/10kb				10kb/2kb				Gene			
Population	**SNP**	**π**	**θ**	**Tajima's D**	**SNP**	**π**	**θ**	**Tajima's D**	**SNP**	**π**	**θ**	**Tajima's D**
Ancestral wild	1998	0.0057	0.0040	0.0785	208	0.0057	0.0040	0.0789	59	0.0046	0.0031	0.0833
Mesoamerican												
Wild	1749	0.0060	0.0040	0.0852	182	0.0061	0.0041	0.0836	53	0.0049	0.0032	0.0771
All landraces	1900	0.0050	0.0037	0.0382	198	0.0050	0.0037	0.0364	56	0.0039	0.0028	0.0249
Mexican landraces	1852	0.0049	0.0039	0.0418	193	0.0050	0.0039	0.0397	55	0.0038	0.0030	0.0316
Mexican sub population 1	1192	0.0035	0.0029	0.0283	124	0.0035	0.0029	0.0255	35	0.0028	0.0022	0.0205
Mexican sub population 2	1149	0.0044	0.0039	0.0281	120	0.0044	0.0039	0.0260	36	0.0036	0.0032	0.0208
Mexican sub population 3	991	0.0027	0.0024	0.0133	103	0.0027	0.0024	0.0104	30	0.0021	0.0019	0.0030
Central American sub population	1378	0.0047	0.0037	0.0459	143	0.0047	0.0037	0.0435	42	0.0037	0.0029	0.0348
Andean												
Wild	555	0.0014	0.0013	0.0067	58	0.0014	0.0013	0.0056	16	0.0010	0.0010	-0.0003
All landraces	618	0.0017	0.0015	-0.0484	64	0.0017	0.0015	-0.0471	20	0.0015	0.0013	-0.1132
Andean sub population 1	273	0.0011	0.0009	0.0203	29	0.0011	0.0009	0.0195	8	0.0009	0.0007	0.0222
Andean sub population 2	507	0.0016	0.0013	0.0171	53	0.0028	0.0014	0.0136	16	0.0014	0.0012	0.0163

Supplementary Table 15. Window or gene based summary of population genomics statistics for common bean averaged over two window sizes and individual genes.

Nature Genetics: doi:10.1038/ng.3008

Supplementary Table 16. Mesoamerican domestication candidates (see separate Excel file)
Supplementary Table 17. Andean domestication candidates (see separate Excel file)

Nature Genetics: doi:10.1038/ng.3008

Comparison	Upper 90% $\pi_{wild}/\pi_{landrace}$	Upper 90% F_{ST}
10kb/2kb sliding window		
Mesoamerica wild vs. landrace		
Mesoamerica	2.5596	0.3806
Andean wild vs. Andean landraces	2.7214	0.3304
Genes		
Mesoamerica wild vs. landrace		
Mesoamerica	4.0510	0.4613
Andean wild vs. Andean landraces	2.9512	0.3103

Supplementary Table 18. Pi-ratio and Fst cutoff values to identify selection.

Nature Genetics: doi:10.1038/ng.3008

Gene Model	Seed weight symbol	Chrom	Start	End
Phvul.001G000500	CDLB1	Chr01	144,309	146,169
Phvul.001G003700	EXPO10	Chr01	341,806	344,139
Phvul.001G007800	LOG1	Chr01	616,683	620,819
Phvul.001G017100	DA1	Chr01	1,429,648	1,435,650
Phvul.001G032200	KNAT1	Chr01	3,078,925	3,084,634
Phvul.001G037400	LOG1	Chr01	3,601,073	3,604,290
Phvul.001G038800	CKX7	Chr01	3,860,546	3,865,008
Phvul.001G043600	AHK5	Chr01	4,529,713	4,537,672
Phvul.001G066000	GA20OX1	Chr01	8,381,806	8,385,990
Phvul.001G125800	ARR9_ATRR3	Chr01	35,340,847	35,342,926
Phvul.001G128800	CKX1_CKX5_CKX6	Chr01	36,632,356	36,635,291
Phvul.001G149400	IPT3_IPT5	Chr01	40,282,825	40,283,742
Phvul.001G166700	ARR24	Chr01	42,826,352	42,827,545
Phvul.001G168500	ARR24	Chr01	43,093,331	43,094,123
Phvul.001G177400	LOG1	Chr01	44,084,869	44,089,144
Phvul.001G181600	EXPO10	Chr01	44,623,426	44,625,593
Phvul.001G194400	LOG1	Chr01	46,037,896	46,042,703
Phvul.001G204900	WEE1	Chr01	46,982,597	46,985,854
Phvul.001G219700	EXPO10	Chr01	48,217,860	48,218,892
Phvul.001G232600	EXPO10	Chr01	49,338,402	49,339,807
Phvul.001G261500	KLU	Chr01	51,618,070	51,619,993
Phvul.002G007600	CLV1	Chr02	878,645	882,943
Phvul.002G024900	DA1	Chr02	2,660,954	2,668,059
Phvul.002G029500	DDM1	Chr02	3,063,228	3,069,240
Phvul.002G029700	DWF4	Chr02	3,098,498	3,103,000
Phvul.002G083600	EXPO10	Chr02	12,900,620	12,902,739
Phvul.002G090900	EIF-5A	Chr02	15,370,102	15,371,825
Phvul.002G107100	ATHK1	Chr02	21,585,247	21,592,897
Phvul.002G152900	EXPO10	Chr02	29,369,190	29,370,828
Phvul.002G169600	SH/SHB1	Chr02	31,271,901	31,279,943
Phvul.002G169700	SH/SHB1	Chr02	31,290,312	31,295,950
Phvul.002G173000	AHK2_AHK3_AHK4	Chr02	32,130,970	32,138,797
Phvul.002G191500	MSI1	Chr02	34,804,672	34,807,990
Phvul.002G202100	CDLB1	Chr02	36,183,223	36,191,730
Phvul.002G246800	REV	Chr02	41,323,900	41,330,159
Phvul.002G282200	ARF2	Chr02	44,603,605	44,608,648
Phvul.002G285000	HSD1	Chr02	44,850,979	44,853,798
Phvul.002G324900	AHK2_AHK3_AHK4	Chr02	48,341,049	48,349,166
Phvul.003G015500	AHK2_AHK3_AHK4	Chr03	1,411,225	1,417,895
Phvul.003G041200	KLU_EOD3	Chr03	4,582,905	4,584,971
Phvul.003G093100	IPT3_IPT5_IPT7	Chr03	19,179,812	19,181,667
Phvul.003G099000	AHP1_AHP3_AHP5	Chr03	24,084,486	24,086,087
Phvul.003G110100	ARR1_ARR2	Chr03	27,714,817	27,718,947
Phvul.003G136400	CKX2_CKX3_CKX5	Chr03	32,803,946	32,807,861
Phvul.003G136500	CKX3	Chr03	32,819,603	32,824,801
Phvul.003G171500	AVP	Chr03	38,248,747	38,253,381

Nature Genetics: doi:10.1038/ng.3008

Phvul.003G183100	KLU_EOD3	Chr03	39,501,091	39,503,000
Phvul.003G187500	EIF-5A	Chr03	39,959,123	39,960,634
Phvul.003G196300	ARR3	Chr03	40,904,491	40,906,528
Phvul.003G196500	ARR3_ARR15	Chr03	40,921,022	40,923,192
Phvul.003G196600	DEL1	Chr03	40,933,175	40,936,648
Phvul.003G213800	EXPO10	Chr03	42,940,817	42,942,083
Phvul.003G253100	DWF4	Chr03	48,110,623	48,114,757
Phvul.003G264600	ATHK1	Chr03	49,180,437	49,187,253
Phvul.004G028800	GASA4	Chr04	3,121,597	3,124,051
Phvul.004G030500	EXPO10	Chr04	3,354,742	3,356,837
Phvul.004G064600	CYP735A1_CYP735A2	Chr04	8,973,882	8,980,231
Phvul.004G123600	GA20OX1	Chr04	39,699,554	39,701,390
Phvul.004G126100	ERL1_ERL2	Chr04	40,138,267	40,146,076
Phvul.004G133200	MET1	Chr04	41,050,065	41,057,146
Phvul.005G022700	LOG1	Chr05	2,008,896	2,010,207
Phvul.005G027100	ARR9_ATRR3	Chr05	2,488,729	2,490,502
Phvul.005G034000	CKX1_CKX5_CKX6	Chr05	3,178,703	3,181,535
Phvul.005G055400	NAC1	Chr05	7,254,487	7,258,898
Phvul.005G091500	FIE/FIS3	Chr05	26,314,512	26,318,717
Phvul.005G109300	ENT3_ENT4_ENT6_ENT7	Chr05	31,905,455	31,909,909
Phvul.005G134000	LOG1	Chr05	36,088,586	36,092,426
Phvul.005G144500	EXPO10	Chr05	37,308,017	37,309,279
Phvul.005G166900	REV	Chr05	39,178,449	39,185,837
Phvul.005G178200	AHP6	Chr05	40,070,379	40,071,717
Phvul.006G029000	CLV1	Chr06	12,372,351	12,376,922
Phvul.006G077200	EXPO10	Chr06	19,594,514	19,596,982
Phvul.006G086800	EXPO10	Chr06	20,544,891	20,546,039
Phvul.006G103700	AN3	Chr06	22,003,259	22,007,707
Phvul.006G122800	CDLB1	Chr06	23,818,584	23,826,499
Phvul.006G128600	REV	Chr06	24,311,622	24,317,626
Phvul.006G154200	IPT5_IPT7	Chr06	26,718,306	26,720,084
Phvul.006G159300	AHP1_AHP2_AHP3_AHP5 -AHP6	Chr06	27,127,859	27,133,631
Phvul.006G193100	ENT3_ENT4_ENT6_ENT7	Chr06	30,034,563	30,042,648
Phvul.006G193300	ENT3_ENT4_ENT6_ENT7	Chr06	30,060,498	30,062,676
Phvul.006G193400	ENT3_ENT4_ENT6_ENT7	Chr06	30,064,924	30,067,226
Phvul.007G028100	IPT1_IPT6_IPT8	Chr07	2,165,646	2,167,688
Phvul.007G064800	GA20OX1	Chr07	5,714,864	5,716,900
Phvul.007G148800	LOG1	Chr07	36,706,872	36,710,937
Phvul.007G166700	FIE/FIS3	Chr07	39,848,075	39,854,258
Phvul.007G167900	AHP1_AHP2_AHP3_AHP5	Chr07	40,027,937	40,032,078
Phvul.007G170100	IPT3_IPT5	Chr07	40,285,383	40,286,351
Phvul.007G183200	AHP1_AHP2_AHP3_AHP5	Chr07	41,941,397	41,943,008
Phvul.007G189200	AN3	Chr07	42,549,946	42,553,336
Phvul.007G207600	EXPO10	Chr07	44,644,205	44,645,485
Phvul.007G269400	LOG1	Chr07	50,766,672	50,770,415
Phvul.007G269500	E2F3	Chr07	50,780,968	50,785,786
Phvul.008G005600	CYP735A1_CYP735A2	Chr08	615,053	619,432
Phvul.008G034700	EXPO10	Chr08	2,903,260	2,905,172

Nature Genetics: doi:10.1038/ng.3008

Phvul.008G037500	EXPO10	Chr08	3,131,412	3,133,124
Phvul.008G038300	SH/SHB1	Chr08	3,234,893	3,240,176
Phvul.008G041200	GASA4	Chr08	3,480,877	3,482,421
Phvul.008G120700	EXPO10	Chr08	15,661,050	15,664,385
Phvul.008G160500	ATRR3	Chr08	41,207,395	41,211,052
Phvul.008G229800	DA1	Chr08	54,461,684	54,466,341
Phvul.008G232200	EXPO10	Chr08	54,674,682	54,676,674
Phvul.008G240800	EXPO10	Chr08	55,529,938	55,532,235
Phvul.008G248000	EXPO10	Chr08	56,264,486	56,266,809
Phvul.008G253500	CDLB1	Chr08	56,750,663	56,754,010
Phvul.008G285800	AHK2	Chr08	59,078,551	59,088,594
Phvul.009G016000	LOG1	Chr09	2,660,091	2,663,107
Phvul.009G019000	EXPO10	Chr09	3,497,612	3,499,649
Phvul.009G034400	PUP1_PUP2	Chr09	7,386,248	7,387,735
Phvul.009G043400	ARR5_ARR16_ARR17	Chr09	8,461,349	8,462,913
Phvul.009G060200	CKX1_CKX3_CKX5_CKX6	Chr09	10,719,497	10,726,239
Phvul.009G078800	LOG1	Chr09	12,794,282	12,797,245
Phvul.009G081800	CKX7	Chr09	13,074,754	13,078,870
Phvul.009G109700	MAX4	Chr09	16,502,002	16,504,992
Phvul.009G110500	REV	Chr09	16,589,773	16,595,618
Phvul.009G131500	GA20OX1	Chr09	19,423,003	19,426,332
Phvul.009G138500	BRI1/DWF2_BRI1_EMS1_	Chr09	20,367,117	20,370,855
Phvul.009G142800	EXPO10	Chr09	20,892,482	20,894,300
Phvul.009G155400	GA20OX1	Chr09	22,617,152	22,620,144
Phvul.009G161900	ARF2	Chr09	23,557,877	23,563,227
Phvul.009G182500	DEL1	Chr09	26,885,942	26,892,635
Phvul.009G182800	ARR7	Chr09	26,960,006	26,962,365
Phvul.009G184600	EIF-5A	Chr09	27,211,707	27,214,154
Phvul.009G186400	EXPO10	Chr09	27,567,879	27,570,714
Phvul.009G187400	GASA4	Chr09	27,698,675	27,700,213
Phvul.009G231700	CKX3_CKX5	Chr09	34,182,344	34,186,263
Phvul.009G231800	CKX3	Chr09	34,223,306	34,229,917
Phvul.009G253200	ARR1_ARR2	Chr09	36,593,953	36,598,004
Phvul.010G010200	EXPO10	Chr10	1,596,184	1,599,156
Phvul.010G087500	GA20OX1	Chr10	32,648,913	32,651,094
Phvul.010G117100	KLU_EOD3	Chr10	38,413,620	38,415,853
Phvul.010G146200	REV	Chr10	41,737,278	41,745,149
Phvul.011G013500	ENT1	Chr11	1,029,308	1,031,750
Phvul.011G014000	CKX1_CKX5_CKX6	Chr11	1,092,381	1,094,971
Phvul.011G031700	DWF1	Chr11	2,752,538	2,755,643
Phvul.011G035800	MSI1	Chr11	3,137,684	3,140,779
Phvul.011G063800	EXPO10	Chr11	5,535,300	5,537,331
Phvul.011G079800	REV	Chr11	7,419,065	7,425,525
Phvul.011G080600	LOG1	Chr11	7,544,162	7,548,685
Phvul.011G097700	E2F3	Chr11	10,137,208	10,143,109
Phvul.011G110200	ENT3_ENT4_ENT6_ENT7	Chr11	14,190,043	14,193,584

Supplementary Table 19. Candidate common bean seed weight genes.

Nature Genetics: doi:10.1038/ng.3008

				Arabidopsis thaliana			
Gene model	**Chrom**	**Start**	**End**	**best hit**	**gene symbol**	**gene description**	**Distance GWAS SNP peak**
Phvul.001G261500	1	51,618,070	51,619,993	AT1G13710	CYP78A5, KLU	cytochrome P450, family 78, subfamily A, polypeptide 5	5,223
Phvul.003G099000	3	24,084,486	24,086,087	AT3G21510	AHP1	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	42,841
Phvul.003G196500	3	40,921,022	40,923,192	AT1G74890	ARR5, ATRR2, IBC6, RR5	response regulator 5	427
Phvul.003G196600	3	40,933,175	40,936,648	AT3G48160	DEL1, E2L3, E2FE	DP-E2F-like 1	0
Phvul.003G253100	3	48,110,623	48,114,757	AT3G50660	DWF4, CYP90B1, CLM, SNP2, SAV1, PSC1	Cytochrome P450 superfamily protein	79,755
Phvul.003G264600	3	49,180,437	49,187,253	AT2G17820	ATHK1, AHK1, HK1	histidine kinase 1	31,991
Phvul.004G064600	4	8,973,882	8,980,231	AT5G38450	CYP735A1	cytochrome P450, family 735, subfamily A, polypeptide 1	32,221
Phvul.006G077200	6	19,594,514	19,596,982	AT2G40610	ATEXPA8, EXP8, ATEXP8, ATHEXP ALPHA 1.11, EXPA8	0	16,867
Phvul.006G159300	7	27,127,859	27,133,631	AT3G21510	AHP1	histidine-containing phosphotransmitter 1	67,893
Phvul.007G166700	7	39,848,075	39,854,258	AT3G20740	FIE, FIS3, FIE1	Transducin/WD40 repeat-like superfamily protein	57,051
Phvul.008G120700	8	15,661,050	15,664,385	AT2G37640	ATEXPA4, ATEXP4, ATHEXP ALPHA 1.6, EXPA4	expansin A4	44,202
Phvul.010G010200	10	1,596,184	1,599,156	AT2G39700	ATEXPA4, ATEXP4, ATHEXP ALPHA 1.6, EXPA4	expansin A4	17,454

Nature Genetics: doi:10.1038/ng.3008

Phvul.011G013500	11	1,029,308	1,031,750	AT1G70330	ENT1,AT, ENT1	WPP domain protein 2	1,212
Phvul.011G014000	11	1,092,381	1,094,971	AT3G63440	ATCKX6, CKX6, ATCKX7	sulfur E2	59,412
Phvul.011G035800	11	3,137,684	3,140,779	AT2G16780	MSI2, MSI02, NFC02, NFC2	Transducin family protein / WD-40 repeat family protein	0

Supplementary Table 20. Mesoamerican seed weight improvement candidate genes.

Nature Genetics: doi:10.1038/ng.3008

Gene model	Chrom	Gene start	Gene end	MA selected gene block assignment	Best *A. thaliana* hit	Top *A. thaliana* hit symbol	Top *A. thaliana* hit description	Distance to SNP
Phvul.001G258300	1	51,408,257	51,411,015	95	AT1G67700		unknown protein	9,590
Phvul.001G258400	1	51,413,288	51,418,655	95	AT3G26020		Protein phosphatase 2A regulatory B subunit family protein	14,621
Phvul.001G260800	1	51,580,507	51,583,778	None	AT1G67440	emb1688	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	34,430
Phvul.002G193900	2	35,052,110	35,052,559	None				9,021
Phvul.003G035600	3	3,571,816	3,577,791	238	AT1G13380		Protein of unknown function (DUF1218)	33,863
Phvul.003G050900	3	6,241,388	6,253,146	245	AT2G04160	AIR3	Subtilisin-like serine endopeptidase family protein	0
Phvul.003G104100	3	25,801,140	25,806,833	267	AT1G09040		non-ATPase subunit 9	21,790
Phvul.003G124100	3	30,403,201	30,405,704	269	AT5G28050		Cytidine/deoxycytidylate deaminase family protein	21,166
Phvul.003G124900	3	30,542,597	30,550,021	270	AT5G17250		Alkaline-phosphatase-like family protein	2,573
Phvul.003G144500	3	34,150,168	34,153,951	275	AT5G57390	AIL5, CHO1, EMK	AINTEGUMENTA-like 5	42,105
Phvul.003G196800	3	40,951,232	40,951,474	293	AT3G48180		Plant protein of unknown function (DUF869)	9,239
Phvul.003G264600	3	49,180,437	49,187,253	323	AT2G17820	ATHK1, AHK1, HK1	histidine kinase 1	31,930
Phvul.003G264700	3	49,189,215	49,192,451	323	AT5G66140	PAD2	vacuolar ATP synthase subunit C (VATC) / V-ATPase C subunit / vacuolar proton pump C subunit (DET3)	35,363
Phvul.003G265400	3	49,227,814	49,229,673	323	AT5G04780		Pentatricopeptide repeat (PPR) superfamily protein	8,570
Phvul.004G057500	4	7,657,458	7,661,986	347	AT3G27320		alpha/beta-Hydrolases superfamily protein	35,580
Phvul.004G066500	4	9,460,848	9,462,568	None	AT5G33370		GDSL-like Lipase/Acylhydrolase superfamily protein	30,307
Phvul.006G070000	6	18,939,698	18,940,811	496	AT5G58580	ATL2, TL2	TOXICOS EN LEVADURA 2	13,650
Phvul.006G070100	6	18,945,216	18,945,869	496	AT3G05200			8,592
Phvul.007G065600	7	5,847,170	5,851,242	538	AT5G62165	AGL42	AGAMOUS-like 42	19,861
Phvul.007G065800	7	5,858,793	5,860,428	538	AT5G51890		Peroxidase superfamily protein	10,675
Phvul.007G066000	7	5,869,640	5,872,158	538	AT4G38010		Pentatricopeptide repeat (PPR-like) superfamily protein	0
Phvul.007G066800	7	5,976,854	5,979,050	None	AT5G51940	NRPB6A, NRPD6A, NRPE6A	RNA polymerase Rpb6	15,141
Phvul.007G066900	7	5,993,698	5,997,594	None	AT2G45750		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	31,985
Phvul.007G071100	7	6,408,764	6,409,930	None	AT5G52390		unknown protein	36,657
Phvul.007G075800	7	6,962,320	6,964,291	542	AT5G52870		unknown protein	13,076
Phvul.007G075900	7	6,978,392	6,980,784	542	AT4G23630		Reticulon family protein	12,355
Phvul.007G076300	7	7,016,455	7,024,238	542	AT4G28000		P-loop containing nucleoside triphosphate hydrolases superfamily protein	0
Phvul.007G094000	7	9,664,594	9,665,620	546				41,860
Phvul.007G094200	7	9,696,188	9,705,030	546	AT1G48850	EMB1144	chorismate synthase, putative / 5-enolpyruvylshikimate-3-phosphate phospholyase, putative	46,778
Phvul.007G094300	7	9,724,193	9,729,678	546	AT2G39220	PLP6, PLA IIB	PATATIN-like protein 6	22,130
Phvul.007G094400	7	9,770,994	9,772,378	546	AT5G19290		alpha/beta-Hydrolases superfamily protein	19,186
Phvul.007G095000	7	9,869,289	9,872,915	546	AT4G30080	ARF16	auxin response factor 16	29,711
Phvul.007G095100	7	9,882,588	9,884,606	546	ATMG00300			18,020
Phvul.007G095300	7	9,922,891	9,926,868	546	AT3G54810	GATA9	GATA transcription factor 9	20,265
Phvul.007G095600	7	9,983,235	9,985,678	546	AT5G03250		Phototropic-responsive NPH3 family protein	39,122
Phvul.007G095700	7	9,987,742	9,988,152	546	AT4G02210			36,648
Phvul.007G095800	7	9,992,495	9,994,474	546				30,326
Phvul.007G095900	7	9,996,844	9,998,891	546	AT4G14145		unknown protein	25,909
Phvul.007G097100	7	10,248,037	10,254,281	546	AT3G10360	APUM2, PUM2		34,621
Phvul.007G097200	7	10,278,772	10,289,556	546	AT2G39130		Transmembrane amino acid transporter family protein	0
Phvul.007G097400	7	10,374,618	10,378,029	546				10,474
Phvul.007G097500	7	10,403,778	10,406,331	546	AT4G02550			39,634
Phvul.007G098700	7	10,512,291	10,516,538	546	AT3G54850	ATPUB14, PUB14	plant U-box 14	8,126
Phvul.007G098800	7	10,517,794	10,543,382	546	AT3G10380	SEC8, ATSEC8	subunit of exocyst complex 8	13,629
Phvul.007G098900	7	10,543,501	10,547,486	546	AT2G39140	SVR1	pseudouridine synthase family protein	791
Phvul.007G099100	7	10,588,199	10,591,587	546	AT2G39170		Galactose oxidase/kelch repeat superfamily protein	10,025
Phvul.007G099300	7	10,628,850	10,631,945	546	AT3G10405		unknown protein	27,238
Phvul.007G099500	7	10,638,977	10,646,055	546	AT3G54880		unknown protein	37,365
Phvul.007G100800	7	10,964,374	10,965,093	None				15,225
Phvul.007G101400	7	11,174,527	11,175,292	550	AT1G68765			29,497
Phvul.007G101600	7	11,247,272	11,249,964	550	AT3G25670		RNI-like superfamily protein	42,483
Phvul.007G107600	7	13,015,723	13,017,441	None	AT5G03120			14,372
Phvul.007G108100	7	13,277,836	13,278,510	556			HR-like lesion-inducing protein-related	4,704
Phvul.007G109200	7	13,591,085	13,594,278	None	AT2G30580	DRIP2	DREB2A-interacting protein 2	25,940
Phvul.007G119600	7	19,539,118	19,541,066	575	AT2G02240	MEE66	Transducin family protein / WD-40 repeat family protein	37,574
Phvul.007G121500	7	21,620,803	21,622,206	580	AT1G08650	PPCK1, ATPPCK1	cation/H+ exchanger 20	25,431
Phvul.007G123000	7	23,296,799	23,300,482	None	AT1G54450			44,015
Phvul.007G166700	7	39,848,075	39,854,258	595	AT3G20740	FIE, FIS3, FIE1	Transducin/WD40 repeat-like superfamily protein	9,246
Phvul.007G166900	7	39,863,504	39,867,874	595	AT4G03110	AtRBP-DR1, RBP-DR1	RNA-binding protein-defense related 1	49,780
Phvul.007G171000	7	40,345,396	40,349,737	None	AT1G61750			46,340
Phvul.008G062800	8	5,704,198	5,704,599	None	AT5G12060		Plant self-incompatibility protein S1 family	2,234
Phvul.008G100300	8	10,901,891	10,903,411	None	AT2G41475			34,281
Phvul.008G113700	8	13,662,384	13,663,511	None	AT3G09270	ATGSTU8, GSTU8	glutathione S-transferase TAU 8	39,152
Phvul.008G130300	8	20,089,563	20,091,716	660	AT1G65450		HXXXD-type acyl-transferase family protein	16,963
Phvul.008G130500	8	20,108,398	20,113,506	660	AT5G48660		B-cell receptor-associated protein 31-like	0
Phvul.008G130600	8	20,139,504	20,145,813	660	AT3G25070	RIN4	RPM1 interacting protein 4	30,825
Phvul.008G130700	8	20,149,301	20,151,400	660	AT3G25100	CDC45	cell division cycle 45	40,622
Phvul.008G141900	8	25,473,533	25,473,985	668			Nucleic acid-binding, OB-fold-like protein	43,018
Phvul.008G168000	8	43,530,648	43,537,164	675	AT1G77760	NIA1, GNR1, NR1	nitrate reductase 1	0
Phvul.009G204800	9	30,290,454	30,293,780	798	AT5G10840		Endomembrane protein 70 protein family	0
Phvul.009G223700	9	33,110,310	33,111,518	808	AT4G22600		ARM repeat superfamily protein	23,096
Phvul.009G234200	9	34,533,509	34,536,079	815	AT5G57090	PIN7, ATPIN7	Auxin efflux carrier family protein	46,220
Phvul.010G101800	10	35,914,907	35,916,655	None	AT4G34138	UGT73B1	UDP-glucosyl transferase 73B1	0
Phvul.010G102300	10	35,938,015	35,938,383	None	AT5G63470	NF-YC4	nuclear factor Y, subunit C4	22,480
Phvul.011G037000	11	3,224,391	3,224,860	878	AT4G38840		SAUR-like auxin-responsive protein family	32,254

Supplementary Table 21 Mesoamerican domestication candidates within 50kb of GWAS peak.

Nature Genetics: doi:10.1038/ng.3008

Gene model	Chrom	Start	End	A. thaliana best hit	A. thaliana gene symbol	A. thaliana gene description	Distance to GWAS SNP peak
Phvul.001G261500	Chr01	51,618,070	51,619,993	AT1G13710	CYP78A5, KLU	cytochrome P450, family 78, subfamily A, polypeptide 5	5,223
Phvul.003G099000	Chr03	24,084,486	24,086,087	AT3G21510	AHP1	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein	42,841
Phvul.003G196500	Chr03	40,921,022	40,923,192	AT1G74890	ARR5, ATRR2, IBC6, RR5	response regulator 5	427
Phvul.003G196600	Chr03	40,933,175	40,936,648	AT3G48160	DEL1, E2L3, E2FE	DP-E2F-like 1	0
Phvul.003G253100	Chr03	48,110,623	48,114,757	AT3G50660	DWF4, CYP90B1, CLM, SNP2, SAV1, PSC1	Cytochrome P450 superfamily protein	79,755
Phvul.003G264600	Chr03	49,180,437	49,187,253	AT2G17820	ATHK1, AHK1, HK1	histidine kinase 1	31,991
Phvul.004G064600	Chr04	8,973,882	8,980,231	AT5G38450	CYP735A1	cytochrome P450, family 735, subfamily A, polypeptide 1	32,221
Phvul.006G077200	Chr06	19,594,514	19,596,982	AT2G40610	ATEXPA8, EXP8, ATEXP8, ATHEXP ALPHA 1.11, EXPA8	0	16,867
Phvul.006G159300	Chr07	27,127,859	27,133,631	AT3G21510	AHP1	histidine-containing phosphotransmitter 1	67,893
Phvul.007G166700	Chr07	39,848,075	39,854,258	AT3G20740	FIE, FIS3, FIE1	Transducin/WD40 repeat-like superfamily protein	57,051
Phvul.008G120700	Chr08	15,661,050	15,664,385	AT2G37640	ATEXPA4, ATEXP4, ATHEXP ALPHA 1.6, EXPA4	expansin A4	44,202
Phvul.010G010200	Chr10	1,596,184	1,599,156	AT2G39700	ATEXPA4, ATEXP4, ATHEXP ALPHA 1.6, EXPA4	expansin A4	17,454
Phvul.011G013500	Chr11	1,029,308	1,031,750	AT1G70330	ENT1,AT, ENT1	WPP domain protein 2	1,212
Phvul.011G014000	Chr11	1,092,381	1,094,971	AT3G63440	ATCKX6, CKX6, ATCKX7	sulfur E2	59,412
Phvul.011G035800	Chr11	3,137,684	3,140,779	AT2G16780	MSI2, MSI02, NFC02, NFC2	Transducin family protein / WD-40 repeat family protein	0

Supplementary Table 22. Mesoamerican seed weight improvement candidate genes.

Nature Genetics: doi:10.1038/ng.3008

B. Supplementary Note

Outline

Nature Genetics: doi:10.1038/ng.3008

1 Sequencing, Assembly, and Annotation

1.1 Accession numbers

Version 1.0 assembly - Assembly and annotation is available from http://www.phytozome.net/commonbean.php and is deposited in Genbank under accession ANNZ01000000.

454 Shotgun and Pairs: SRX012337-SRX012348, SRX028889-SRX028890, SRX028894-SRX028898, SRX028915-SRX028920, SRX028964-SRX028978, SRX062194-SRX062216, SRX273310-SRX273311
BAC END Sequence - PV_A: EI415689-EI504705; PV_B, PV_C: JY504315-JY663793
Fosmid End Sequence – JY665079-JY879798, JY893769-JY972748.
Illumina Whole Genome Shotgun: SRX273308-SRX273309

1.2 Pseudomolecule Chromosome Construction

The combination of the available genetic maps (7,015 SNP and 261 SSR markers for a total of 7,276) as well as 25 framework markers and *Glycine max* synteny were used to identify false joins in the initial assembly. Scaffolds were broken if they contained a putative false join coincident with an area of low BAC/fosmid coverage. A total of 71 breaks were identified and broken, resulting in 1,698 scaffolds in the broken assembly. The optimal order and orientation of the broken scaffolds was obtained using markers and *G. max* synteny. Due to the high-resolution of the genetic map (7,015 markers in the 267-individual primary mapping population) and the large size of the assembled scaffolds, the pseudomolecule assemblies were well constructed before use of synteny. Nevertheless, genetic map data alone was not able to give precise placements or orderings of scaffolds within the recombination-poor pericentromeric regions. Additional refinements to the 11 pseudomolecule chromosomes were made based on synteny with soybean (*Glycine max*). Approximately 22% (52/240) of the initial marker-based scaffold ordering was locally modified based on *G. max* synteny with (usually within a 1 cM range, within the pericentromeres); and 17% (41/240) of the orientations were changed. Almost all such order/orientations and synteny changes were made within the Phaseolus pericentromeric regions, where there is virtually no genetic recombination. Significant telomeric sequence was identified using the TTTAGGG repeat, and care was taken to make sure that it was properly oriented in the production assembly. BAC/Fosmid paired end link support was also used to order and orient the scaffolds composing the pseudomolecule chromosomes. A total of 248 joins were made on 259 scaffolds to form the final assembly containing 11 chromosomes capturing 514.8 Mb (98.8%) of the assembled sequence. Each join is sized with 10,000 Ns. After screening for contaminant, there were 697 additional scaffolds that did contain a marker alignment and could not be localized using *G. max* synteny, and they are included as part of the release assembly. The final assembly contains 708 scaffolds (41,391 contigs) with a contig L50 of 39.5 kb and a scaffold L50 of 50.4 Mb. Plots of the marker placements for the 11 chromosomes are shown in Supplementary Figs. 1-11.

1.3 Screening and Final Assembly Release

Remaining scaffolds were classified into bins depending on sequence content. Contamination was identified using megablast against Genbank NR and blastp using a set

Nature Genetics: doi:10.1038/ng.3008

of known microbial proteins. Additional scaffolds were classified as mitochondrion (8 scaffolds, 18.1 Kb), chloroplast (12 scaffolds, 453.5 Kb), unanchored rDNA (6 scaffolds, 158.1 Kb), prokaryote (1 scaffold, 44.8 Kb), unanchored retrotransposons (28 scaffolds, 65.1 Kb), repetitive (>95% masked with 24mers that occur more than 4 times in the genome) (160 scaffolds, 1.4 Mb). We also removed 527 scaffolds that were less than 1 kb in sequence length (total of 390.4 Kb). Resulting final statistics are shown in Supplementary Table 3.

1.4 Assessment of Assembly Accuracy

A set of 8 random BAC clones totaling 1.12 Mb were sequenced in order to assess the completeness of the genic regions. A low rate of base pair mismatch and indel bases (combined <0.5%) was indicated in the comparison of the 8 BAC clones and the assembly, with the main discrepancies in the clones being minor gaps (2-5 Kb). A representative example of one of these BAC clones is given in Supplementary Fig. 12 (all dot plots were generated using Gepard (Krumsiek and Rattei 2007)). The overall nonmatching bp rate (not including gap bases) in this group of clones is 0.13% (1,414 bp out of a possible 1.03 Mb). A second set of 5 BAC clones aligned to regions of moderate transposon content, with a representative clone given in Supplementary Fig. 13. The third set of BAC clones are ones that place in regions of high transposon content, with an example given in Supplementary Fig. 14. Finally, there are regions where the transposon/repeat content is a confounding factor in the genome assembly process, resulting in these regions not being included in the final assembly. An example of such a clone is given in Supplementary Figure 23. The clone was not located in the final assembly, likely due to the complex repetitive structures in the clone.

Completeness of the euchromatic portion of the genome assembly was assessed using 108,012 *P. vulgaris* EST sequences >400bp obtained from GenBank. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. ESTs were aligned to the release assembly using BLAT (Parameters: -t=dna -q=rna –extendThroughN). Alignments that comprised >=90% base pair identity and >=85% EST coverage were retained. The screened alignments indicate that 102,254 of 108,012 (96.9%) of the ESTs aligned to the assembly. A further 2,146 (2.03%) could be placed at >50% EST coverage, totaling 98.93%. Comparatively few sequences represented artifacts (2,479;2.3%) or were not found (1,133;1.07%). We also aligned 11 rnaSEQ libraries composed of 2x100 bp Illumina reads given in Supplementary Table 4. Reads were aligned using GSNAP (Wu and Nacu,2010) with paramters "-A sam -N 1 -n 6 -w 5000 --nthreads=1 --novelend-splicedist=5000 -K 18 -l 18 --pairmax-rna=5400 --max-mismatches=0.04" as part of the annotation process with an average of 93.2% aligned to the genome sequence (Supplementary Table 4).

2 Centromere and Pericentromeric Analysis

Centromeric positions were identified by BLASTN using centromere tandem repeats CentPv1 and CentPv2 with at least 80% length similarity and 60% identity. For almost all chromosomes CentPv1 was used except for Pv05, Pv06 and Pv11 where CentPv2 was used (Iwata et al. 2013).

To determine the proportion of the genome that falls within pericentromeric

Nature Genetics: doi:10.1038/ng.3008

regions, we compared gene and repeat density and genetic distance versus the physical distance (Supplementary Fig. 16). Genetic distance was measured using 6945 SNP and SSR markers on the assembly, in the Stampede x Redhawk F2 population genetic map with 267 individuals. Repeats density was parsed using Repeatmasker (version 3.3.0; http://www.repeatmasker.org/) with non-default parameters based on a TE custom library constructed for Phaseolus which include 791 repeats composed by 285 Class I elements, 460 Class II elements and 46 unclassified elements in the database (www.phytozome.org). All measures were taken per 1-Mb sliding window at 200-kb intervals; gene counts were taken for gene density, nucleotide counts for repeats density and average of cM between markers in a window were taken for genetic distance. The start and end points on the pericentromeric regions were taken according the cross points in the plots where gene density decreased, repeats density increased and the recombination rate is suppressed or diminished.

3 Repeat Annotation and Analysis

Transposons are the most abundant genetic elements which have broad impacts on genome evolution, gene innovation and regulation, as well as on maintenance of chromosome structure and genomic heterochromatic silencing (Lippman and Martienssen 2004). In addition, transposons also serve as useful tools for insertional mutagenesis and gene isolation (Kumar and Bennetzen 1999; IRGSP 2005). Thus, genome-wide transposon annotation is important for understanding the genome composition and dynamics and the initial step for discovering endogenous active transposons in common bean.

The common bean genome harbors ~45.0% of transposons, which include 39.4% of retrotransposons (Class 1) and 5.5% of DNA transposons (Class 2) (Supplementary Table 8). The Ty3-gypsy retrotransposons are the most plentiful elements which make up about 25.1% of the genome or more than 50% of the total transposons. Ty1-copia retrotransposons account for about 9.4% of the common bean genome. In addition, some LTR retroelements cannot be grouped as their internal regions encode no retrotransposase or only produce tiny proteins, these elements constitute 2.0% of the genome. Long interspersed elements (LINEs) and short interspersed elements (SINEs) comprise 2.9% and 0.03% of the common bean genome. DNA elements are much lower than retroelements in number and fraction, they contribute 5.5% of the common bean genome. Among DNA transposons, the CACTA elements are the most abundant superfamily, these elements constitute 2.7% of the genome. In addition, four superfamilies of DNA elements also were identified which include Harbinger/PIF, hAT, Helitron and MULE. The transposon contents on 11 chromosomes in common bean are different, the chromosomes 10 exhibits the highest transposon content (55.1%) whereas the chromosome 9 has the lowest fraction of transposons (26.8%) which is less than half of chromosome 10 (Supplementary Table 9). The proportions of DNA transposons on 11 chromosomes are similar which range from 5.0% on chromosomes 2 and 7 to 6.1% on chromosome 4. However, the retrotransposon contents greatly vary from 20.8% on chromosome 9 to 48.7% on chromosome 10 and suggesting that the difference of transposon fractions on 11 chromosomes was mainly caused by retrotransposons.

Nature Genetics: doi:10.1038/ng.3008

To gain insight into the dynamics of LTR retrotransposons, the integration times of 2668 full length LTR retroelements were calculated (Supplementary Fig. 1). Most, 75% (2011/2668), of LTR retroelements integrated into common bean within the last 2 million years (MY), although, someancient elements that inserted into the genome more than 10 million year ago (MYA) were also found. Notably, the insertion times of 20% (543/2668) of the elements were less than 0.5 MYA, this result likely suggests that these elements inserted recently and some of them may be still active in the genome.

The insertion dynamics of retroelements on the11 chromosomes vary (Supplementary Fig. 2). More than 84.0% of the complete elements on chromosomes 10 and 11 were inserted less than 2 MYA, however, only 57.0% of the elements on chromosome 9 were integrated within 2 MYA, which is lower than that (65.3% to 78.5%) on other 8 chromosomes.

The 2668 complete LTR retrotransposons were grouped into 165 families including 65 Ty1-copia, 78 Ty3-gypsy and 22 unclassified families according to the described criteria (Wicker et al. 2007). These 165 LTR retrotransposon families contain different numbers of complete retroelements. More than 78% (130/165) of LTR retrotransposon families have less than 10 complete retroelements, however, more than 50 complete elements were found for each of 11 families which contain totally 63% (1690/2668) of the complete elements in common bean genome. It is worth noting that some families show extremely high copy numbers. For example, a small retroement family named pvRetroS2 contains 446 complete elements. Interestingly, the size of pvRetroS2 is only 342 bp and with 122-bp LTR, thus this family may be considered as the terminal-repeat retrotransposons in miniature (TRIM) group (Witte et al. 2001). Other two Ty3-gypsy families, pvRetro31 and pvRetro48, have 364 and 156 complete copies, respectively. To explore the amplification dynamics of different retrotransposon families, the insertion times of 11 families are compared (Supplementary Figure 3). The insertion times of pvRetroS2 elements range from 0 to more than 10 MYA and no obvious amplification peak was found, this suggests that the amplification events of pvRetroS2 retroelements occurred over a long period and these elements may have an ancient origin. However, the other 10 families show dramatic difference in amplification dynamics with pvRetroS2, most elements of these 10 families inserted in the common bean less than 2 MYA. Impressively, more than 44% (163/364) of pvRetro31 elements were inserted less than 0.5 MYA.

Compared to other sequenced plants, the transposon fraction in common bean is larger than that in rice of 35% [3], but is less than 52% in pigeonpea (Varshney et al. 2001) and 59% in soybean (Schmutz et al. 2010), 62% in sorghum (Paterson et al. 2009) and 85% in maize (Schnable et al. 2009). Despite Ty3-gypsy elements are most abundant in these genomes, however, the ratios of Ty3-gypsy to Ty1-copia are different. The ratio is about 2.5:1 ratio in common bean, it is similar to that of 2.4:1 ratio in soybean, but is lower than that in rice (2.8:1) and sorghum (3.71). LINEs contribute 1.0% of maize, rice and pigeonpea genomes, 0.25% of soybean and 0.04% of sorghum. However, nearly 3.0% of common bean genome is comprised of LINEs. The DNA transposon content is 5.5% in common bean, lower than found in rice (12.3%), sorghum (7.5%), maize (8.6%), and soybean (16.5%). Other than rice, CACTA elements are the most abundant among different superfamilies of DNA elements in the sequenced genomes.

In summary, our results indicate that: 1) The common bean genome harbor 45.4% transposons which is similar to that (45%) in human; 2) The common bean genome likely

Nature Genetics: doi:10.1038/ng.3008

have undergone massive amplification of LTR retrotransposons within 2 MYA; 3) 165 LTR retrotransposon families were detected in common bean, the majority of these retrotransposons show low transposition activity.

4 Resistance Gene Analysis

The complete set of NL proteins was identified in a reiterative process. First, an HMM search of the predicted protein sequences of Phaseolus (*Phaseolus vulgaris G19833*; JGI, version 1.0) was done to identify sequences containing NB-ARC domain. The "trusted cutoff" of the NB-ARC domain HMM (PF00931) established by Pfam (Finn et al. 2010) was used as the threshold for detecting NBS domains. This analysis led to the identification of 398 predicted proteins corresponding to 342 annotated genes that encoded homologs of NL proteins. To identify homologs (such as diverse or not being identified as ORFs by the automated annotation) missed in the first step, all the NL predicted protein sequences identified in the first step were used as query to tBLASTn the entire genome. All resulting sequences in the BLAST output (E value < 1e-10) were manually inspected using the Artemis software tool (Rutherford et al. 2000). This procedure identified 34 additional NL genes. A new identifier was created for each missing genes (the last digits are 50).

Domain predictions and manual annotation

NL genes were assessed manually in Artemis software for the presence of TIR (PF01582), NB-ARC (PF00931) and LRR (PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13504 and PF13855) domains with HMMer using trusted cut-off defined in Pfam. Coiled Coil domains were identified using Coils (Lupas et al. 1991) with a 14 amino-acid search window and a 2.9 score cut-off threshold. All this information was imported into the annotation platform Artemis for further manual analysis. We classified sequences with stop codons and/or frameshift as pseudogene.

5 Comparison of Glycine and Phaseolus

The *Glycine max* genome was used as a reference for identification of synteny and for estimates of gene divergence rates between *Glycine* and *Phaseolus*. Synteny blocks within and between *Glycine* and *Phaseolus* were identified by first making blast comparisons of peptide sequences, followed by filtering to top hits per chromosome pair, and then synteny prediction with DAGchainer (Haas et al. 2004). The Ks values for for gene pairs from synteny blocks were calculated, using in-frame CDS alignments, using the codeml program from the PAML package. Mean values per synteny blocks were then taken; histograms of block-mean Ks values are shown in Supplementary Fig. 17.

Syntenic blocks are generally highly collinear with Glycine, except in the pericentromeric regions – where synteny is extenuated due to low gene density. The order and structure of synteny blocks in Glycine versus Phaseolus confirm previous studies on synteny at genetic linkage map level (Galeano et al. 2011; Galeano et al 2009; McClean et al. 2011). For most Phaseolus genes, it is possible to find strongly homologous genes in at least 2 homoeologous chromosomes of Glycine (Main Figure 1), due to the soybean paleotetraploidization (Gill et al. 2009; Schmutz et al. 2010).

Nature Genetics: doi:10.1038/ng.3008

The average numbers of homologous genes per synteny block in the Glycine - Phaseolus and Phaseolus – Phaseolus comparisons are 33 and 14 genes, respectively. Of the Phaseolus genes, 91% (24,861) are contained in synteny blocks with Glycine (via the ~20 Mya speciation), and 57% are in synteny blocks within the Phaseolus – Phaseolus comparison (via the ~58 Mya WGD). Similarly, 86% (46,853) of the total genes in the Glycine genome are in synteny blocks within the Glycine-Glycine comparison (via either the ~10 Mya or the ~58 Mya WGDs), and 96% (46,814) of those Glycine genes are in synteny blocks with Phaseolus (86% of the total genes in the Glycine genome).

Using the modal Ks values from the Ks plots, we determined the likely branch lengths (in Ks units) for the Glycine, Phaseolus, and "shared" portions of an idealized Glycine-Phaseolus gene tree (Supplementary Figure 17). There are three types of paths between leaves (genes) in this tree. Each may be represented as an equation, with the value of the equation being the modal Ks value for that path.

As evident in Ks plots of synteny-block-median Ks values from gene pairs from syntenic regions (Supplementary Figure 17), Phaseolus has evolved faster than soybean since their common ancestor. Assuming that Glycine and Phaseolus separated at 19.2 Mya (7), the Ks rate along the Phaseolus lineage is 0.1625/19.2 Mya = 8.4635 e-9, and the Ks rate along the Glycine lineage is 0.1125/19.2 Mya = 5.8594 e-9. The Phaseolus rate has therefore 1.44 times faster than the Glycine rate, since their common ancestor. Using the sharp Ks peak of 0.125 for the Glycine-Glycine WGD, the estimated time to that palaeotetraploidization would be (0.125/2)/5.859 e-9 = 10.6 Mya (Supplementary Figure 17).

Estimates of the whole-genome duplication (WGD) time range from 45.8 and 57.6 Mya, depending on use of the faster Phaseolus Ks rate or the slower Glycine rate from the common ancestor of Glycine and Phaseolus to the legume WGD episode. This range contains the estimate from Lavin et al. (2005) of 56.5 Mya for the papilionoid radiation, and is similar to the estimates of 44-58 Mya in Schlueter et al. (2004) and Schmutz et al. (2010).

Fractionation and locally duplicated gene clusters

Gene loss and gene retention was identified taking the genes shared and non-shared between *Phaseolus* and *Glycine*. The list of the *Phaseolus* genes retained was used to do a BLASTp analysis against *Glycine* with an E-value ≤1e-10 whit a cutoff of 80% length and 80% identity, to confirm whether they are lost or moved in the *Glycine* genome, and conversely for *Glycine* genes retained versus *Phaseolus*.

To identify locally duplicated genes in *Phaseolus* and *Glycine*, a BLAST comparison between whole chromosomes in *Phaseolus* and whole chromosomes in *Glycine* was parsed, genes similar at E-value ≤1e-10 and clustered within sliding windows of 100 kb, were taken as locally duplicated genes. Over the total of genes in GmPv synteny blocks, 21% (5203/24861) of those genes are locally duplicated in the *Phaseolus* genome and 17% (7849/46814) are locally duplicated in the *Glycine* genome. Furthermore, 20% (5082) of the synteny genes are retained in *Phaseolus* with respect to *Glycine*, and 26% (12269) of the genes are retained in *Glycine* in contrast with *Phaseolus*.

The *Phaseolus* synteny sites, which have copy in at least one homolog in soybean were analyzed per chromosome (Supplementary Table 11), resulting in 1798 *Phaseolus* synteny genes on average per chromosome, having chromosome 2 the highest number of

Nature Genetics: doi:10.1038/ng.3008

synteny sites with 2695, corresponding with the highest number of genes in the genome (3338) and with the major number of ortholog genes in *Glycine*. In the same way, chromosome 10 covers the fewest number of synteny sites (1020), corresponding with the slight number of genes in the genome (1659). Fractionation occurs almost in the same proportion in both copies of the *Glycine* genome, only 21 genes in *Glycine* have a third paralog gene (not included in the table).

Structural organization

The synteny blocks identified for *Phaseolus* – *Glycine*, *Glycine* – *Glycine* recent duplication and *Phaseolus* – *Phaseolus* were taken to make the reference rings in a Circos graph for visualization (Krzywinski et al. 2009) . Homologous genes in *Phaseolus* derivated after speciation are showed with connection lines.

Based on *Phaseolus* data, gene density and repeats density were parsed as described below and recombination rate was parsed dividing the distance in cM between the markers in the genetic map, by the distance in Mb between the markers in the sequence map, taking the midpoint of the location of the markers in the sequence. Sliding windows of 1-Mb at 200-kb intervals was taken and finally the windows with high discrepancies were eliminated.

Polyploidy and fractionation

One effect of polyploidy is fractionation, or loss of genetic material from one or both duplicated chromosomes. Using *Phaseolus* and *Glycine*, we analyzed fractionation from the shared WGD and the more recent WGD unique to *Glycine*. Fractionation occurred in similar proportions in both duplicated copies of the *Glycine* genome (Supplementary Fig. 17). However, based on combined phylogenetic and synteny analyses, we estimate that 9% of the apparent differential gene loss between *Glycine* and *Phaseolus* relative to their shared (pan-legume) duplication is due to expansion of gene clusters in one or the other of the genomes, rather than to selective loss of low-copy (unclustered) genes.

Surprisingly, *Phaseolus* genes occur in locally duplicated clusters at a rate 25% higher than *Glycine* (17.3% in *Glycine* versus 21.5% in *Phaseolus*). Nevertheless, due to the recent WGD in *Glycine*, there are 60% more locally clustered genes in *Glycine* than *Phaseolus*, and the total number of paralogs in *Glycine* is much higher (16,919 in *Glycine* versus 3,197 in *Phaseolus* – or 31% versus 12% of total genes).

6 Historical Population Size Analysis

Divergence of wild Mesoamerica and wild Andean pools

A recent study based on five gene loci from a wide collection of wild common bean samples (Bitocchi et al. 2012) pointed to Mesoamerica as the origin of all common bean varieties existing today. There are two major gene pools for the wild *Phaseolus vulgaris*, wild Mesoamerica and wild Andean, which underwent two independent domestications giving rise to all the major landraces. To investigate the details of the divergence and demographic history of the two wild pools, we make use of the whole genome pooled

Nature Genetics: doi:10.1038/ng.3008

sequencing data (Supplementary Table 12) consisting of 30 individuals within each pool, and make inferences about the demographic parameters by modeling the joint allele frequency spectrum (jAFS) using the package *dadi* version 1.6.3 (Gutenkunst et al. 2009).

To minimize bias in our demographic inference due to selection effects, we used neutral sites which are defined to be at least 5kb away from a gene (as annotated in the gff3 file v1.0) and are not located in the repetitive regions (as defined by Repeatmasker (Smit et al. 1996)). Due to the high selfing rate (~93%) in common bean (Ibarra-Perez et al. 1997), the number of different haplotypes for each pooled sample is close to 30. The data were thus down-sampled to 25 haplotypes for each pool via hypergeometric projection (i.e. random sampling 25 alleles without replacement), from which the joint allele frequency spectrum (jAFS) was derived. As spurious singletons can arise due to sequencing and mapping errors, we excluded sites appearing as singletons in either of the two pools, resulting in a total of 662,835 polymorphic sites for the jAFS.

We investigated and compared different demographic models based on the relative log-likelihoods of the models given the observed site frequency spectrum. No population growth or decline was detected in the ancestral population before the two pools split. Based on this and other observations, we select a model (Main text Figure 1) with constant population size before the divergence of the two pools, and allow an epoch of constant population size for the wild Andean after it split from the wild Mesoamerican population, followed by an exponential growth phase till the present. By contrast, for the wild Mesoamerican population, a single epoch of exponential growth is adequate to describe its post-divergence history. Asymmetric migration rates are assumed in the model (Main text Figure 1).

To make inference of model parameters, we ran dadi simulations with different starting points in an 8-dimensional parameter space, till convergence is achieved. Parameter values for the best fit model are listed in Supplementary Table 13, using a base substitution rate mu=8.46e-9/bp/yr (S. Cannon, unpublished) derived from silent sites. To estimate parameter uncertainties, we divided the genome into 10cM segments and performed 100 bootstraps on the chromosome segments. Confidence intervals were derived based on simulation results for the bootstrapped samples. The results are shown in Supplementary Table 13.

Comparisons between model prediction and observed data are shown in Supplementary Figs. 24 and 25. Supplementary Fig. 24(a) shows the summary statistics of 4 types of mutually exclusive single nucleotide variants, with 80% of all variants accounted for by the wild Mesoamerican pool (MW) alone. By contrast, only 12.5% of the variants are observed exclusively in the wild Andean pool (AW). This great disparity in genetic diversity between the two pools can be explained by the strong population bottleneck in the Andean gene pool and is consistent with the Mesoamerican origin of the common bean (see discussion later). The marginal allele frequency distribution for each of the two pools was shown in Supplementary Figs. 24(b) and Fig. 24(c), respectively, with good agreement between model prediction and data.

The joint allele frequency spectra between the two pools are shown in Supplementary Figure 25. The difference between the model and data is described by Anscombe residuals following dadi (Gutenkunst et al. 2009), and is shown in the lower panel. As can be seen from the lower left panel of Supplementary Fig. 25, the model predicts fewer sites with low-frequency alleles in both pools, and an excess of sites with

Nature Genetics: doi:10.1038/ng.3008

large allele frequency differences between the two pools. These discrepancies may reflect a more complex history of the common bean than captured by the model presented here. For example, the migration rates are more likely to be time-varying than stationary, as the wild Andean population size had changed by a factor of ~60 since its founding population. Another feature unaccounted for by our model is the possible genetic structure within the wild Mesoamerican gene pool (Bitocchi et al. 2012). These and other details may be resolved with additional sequencing beyond the two pooled datasets.

7 Common Bean Domestication Analysis

Development of common bean wild and landrace populations for pooled resequencing.

Initially, 135 wild and 180 landrace genotypes, collected from the full geographic range of *P. vulgaris*, were scored with 22 indel markers (Mafi Modhaddam et al. 2013) distributed throughout the genome. A Bayesian analysis was performed on the genotype data within each of the two groups using the STRUCTURE software (Pritchard et al. *2000a*; Falush et al. *2003*). The linkage ancestry model with correlated allele frequencies was used to analyze the data with a haploid phase setting because common bean is self-fertilizing species. Based on previous experience with a subset of this population (McClean et al. 2012), a total of 20,000 iterations were performed following a burin length of 50,000. In each case, the number of subpopulations ranged from k=2 to k=10 with 10 runs for each subpopulation size. For the wild genotypes, k=2 best fit the data (Evanno et al. 2005). These subpopulations correspond geographically to the wild Mesoamerican and wild Andean gene pools. Because many studies have described further substructure in common bean landraces, k=6 was chosen to further subdivide the landrace genotypes. At k=2, Mesoamerican and Andean landrace subpopulations were defined. At k=3, the Mesoamerican landraces where split into Mexico and Central American subpopulations. At k=4 and k=5, the Mexico subpopulation was further split into three subpopulations. The original Andean subpopulation at k=2 was retained from k=3-5, and at k=6, the southern and northern Andean landrace subpopulations were defined. A genotype was assigned to subpopulation if its subpopulation parentage was >70%. Based on this STRUCTURE analysis, we developed pooled populations for sequencing. From each wild subpopulation, 30 individuals were selected to create wild Mesoamerican and Andean populations for pooled sequencing. All members of each subpopulation were from distinct geographic locations. The average parentage for each genotype within each wild pool was 98%. Similarly, six landraces populations were developed for pooled sequencing (Supplementary Table 12). Average parentage for members in these populations ranged from 90% to 96%. A graphical display of the population membership of the genotypes selected for pooled resequencing is found in Supplementary Fig. 18.

DNA sequencing and SNP identification.

DNA from each of these pooled populations was sequenced to ~4X depth using Illumina technology. Each read was mapped to v1.0 version of the assembled reference genome using BWA (Li and Durbin 2009) with maximum number of hits set to 8. All reads with a quality score less than 25 were discarded. An mpileup file was created for each sequenced

Nature Genetics: doi:10.1038/ng.3008

pool using SAMtools (Li et al. 2009) with the –BA options. VarScan 2.2.10 (Koboldt et al. 2012) utilized the mpileup file for SNP calling with the following parameters: minimum coverage = 5; minimum consensus quality = 25, minimum variant frequency = 0.01. To further reduce SNP call quality, 1) a SNP was discarded if the reference or variant allele was a 'N'; 2) a SNP was discarded if more than one variant allele was observed; and 3) if the variant allele was a single nucleotide indel that position was discarded.

Similar to previous work in chicken and pig (Rubin et al. 2010, 2012), SNP data from several pooled populations were combined. Mesoamerican and Andean landrace population SNP diversity data were created by combining SNP data for each of the appropriate race pools. By pooling the SNP data from these pools, we were able to create datasets representative of the diversity found within the early domestication populations from which landraces were subsequently derived. Additionally, the data from the three Mexican subpopulations were combined to create a single race Mexican landrace pool. The minimum number of reads required for the reference or variant allele was three. The number of SNPs ranged from 8,890,318 for the wild Mesoamerican pool to 1,397,405 for the Peru landrace pool (Supplementary Table 14). Among all wild genotypes, 10,158,326 SNPs were observed while the Mesoamerican landraces contained 9,661,807 SNPs, and all Andean landraces 3,154,648. For all individual and combined pools, the proportion of SNPs found within genes was ~16% indicating that the genes were not disproportionately prone to more (or less) variation.

Population genetics statistics.

Several population genetics statistics were calculated for each 100kb/10kb and 10kb/2kb sliding window, and each gene within each DNA pool. Any window or gene with >50% Ns were excluded, and all statistics were based on the number of non-N nucleotides in the window. Nucleotide diversity (π; Tajima 1983), defined as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population, was calculated using the following formula.

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{i} x_i x_j \pi_{ij}$$

Here, x_i and x_i are the respective frequencies of the i^{th} and j^{th} sequences, π_{ii} is the number of nucleotide differences per nucleotide site between the i^{th} and j^{th} sequences and n is the number of sequences in the sample. The Watterson estimate ($\theta_{w;}$ Watterson 1975), which is an estimation of population mutation rate, was calculated based on the number of segregating sites using the formula

$$\theta w = \frac{S}{a_n}$$

where S is the number of segregating sites and

Nature Genetics: doi:10.1038/ng.3008

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima's D was calculated as described in Tajima (1989). F_{ST}, (Hudson et al 1992) is a measure of population differentiation, estimated from the average pairwise differences between chromosomes in each analysis panel compared to the combined samples as described in The International HapMap Consortium (2005).

$$F_{ST} = 1 - \frac{\sum_j \binom{n_j}{2} \sum_i 2 \frac{n_{ij}}{n_{ij}-1} x_{ij}(1-x_{ij}) / \sum_j \binom{n_j}{2}}{\sum_i 2 \frac{n_i}{n_i-1} x_i (1-x_i)}$$

where x_{ij} is the estimated frequency of the minor allele at SNP i in population j, n_{ij} is the number of genotyped chromosomes at that position, and n_j is the number of chromosomes analyzed in that population. The lack of the j subscript in the denominator indicates that statistics n_i and x_i are calculated across the combined data sets.

The relative diversity level among two pooled samples was compared by a nucleotide diversity (π) ratio between the two pools for each window or gene. For example, the ratio $\pi_{MA\text{-}wild}/\pi_{MA\text{-}landrace}$ measures the relative difference in diversity between the Mesoamerican wild gene pool and the Mesoamerican landrace gene pool. Similarly, F_{ST} (TIHC 2005) was calculated for each window and gene to compare the differentiation between any two pools.

Identifying selected windows and genes and defining sweep windows.

A number of statistical approaches are currently favored when evaluating genome-wide resequencing data to discover genomic regions or genes that are putatively undergoing selection. Divergence approaches use a comparison of nucleotide diversity between an ancestral state and a derived state. These primarily include diversity ratios (Huang et al. 2012; Xu et al. 2012) or reduced heterozygosity (Rubin et al. 2010) among populations. Other studies have used population differentiation methods, such as F_{ST} to identify selected regions (Lam et al. 2010; Turner et al. 2010). Rather than relying on a single statistic, we adopted a strict composite scoring system that combined diversity and differentiation data to identify putative genomic regions or genes under selection. This is similar to the approach applied to silk moth where a reduction in nucleotide diversity and Tajima's D was applied to discover domestication genes (Xia et al. 2009). Here, a 10kb/2kb window or a gene was considered a selection window or domestication candidate gene if it was in the upper 90% of a bootstrap simulation population (n=1000) for the $\pi_{wild}/\pi_{landrace}$ ratio and F_{ST} statistics. The cutoff values for various comparisons can be found in Supplementary Table 18. All 10kb/2kb selection windows within 40kb of each other were merged in a "sweep window". The number of domestication candidates and total genes were calculated for sweep window.

Annotating common bean seed weight/size candidates.

Nature Genetics: doi:10.1038/ng.3008

We identified candidate common bean seed size genes by a blastp analysis using Arabidopsis seed size/weight genes (Van Daele et al. 2012) as a query against a database of the common bean protein sequences. Any common bean gene model hit with 50% identity and 80% coverage that matched 70% of the query length inherited the Arabidopsis seed weight gene name. A total of 141 common bean gene models inherited the seed weight gene name (Supplementary Table 19).

Association Mapping

As part of the USDA Common Bean Coordinated Agricultural Project, a collection of 280 diverse modern common bean varieties from the Middle American gene pool were grown in replicated field trials by the North Dakota State University, Michigan State University, University of Nebraska, and Colorado State University bean breeding programs. Each genotype in the trial was genotyped with 34,799 SNPs. Of these, 10,318 SNPs were from the Illumina Infinium platform used to develop the SNP-based genetic map (see Methods Summary), and 24,481 SNPs were obtained by genotype-by-sequencing (GBS) technology (Elshire et al. 2011). The GBS data was generated by the Institute for Genomic Diversity, Cornell University. Missing data were imputed in fastPHASE 1.3 (Scheet and Stephens 2006). Adjusted means for seed weight data across all locations were calculated using the MIXED procedure in SAS9.3 (SAS 2002) where the genotype was the fixed effect and all other factors were considered as random. A mixed linear model (MLM) controlling for population relatedness was used to conduct the genome wide association study (GWAS). Multiple statistical models were tested, and a mixed model (Yu et al. 2005) that controlled for genotype relatedness and population structure was chosen. An identity-by- state (IBS) kinship matrix [EMMA, (Kang et al. 2008)] was used to control for population relatedness, while two principal components were used to control for population structure. The kinship matrix was calculated using marker loci with pairwise $r^2 > 0.5$. Linkage disequilibrium (r^2) between all marker loci was calculated in Plink (Purcell et al. 2007) using loci with a minor allele frequency (MAF) ≥ 0.05. The EMMA kinship matrix and the GWAS were calculated in the GAPIT package in the R programming language (Lipka et al. 2012), without P3D and compression.

References

Bitocchi, E., L. Nanni, E. Bellucci, M. Rossi, A. Giardini, P.S. Zeuli, G. Logozzo, J. Stougard, P. McClean and G. Attene. 2012. Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Proc. Natl. Acad Sci. 109788-796.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One 6: e19379.

Evanno, G., Regnaut, S., & Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14: 2611-2620.

Falush, D., Stephens, M., & Pritchard, J. K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Nature Genetics: doi:10.1038/ng.3008

Finn, R.D., J. Mistry, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy and A. Bateman. 2010. The Pfam protein families database. Nucleic Acid Res. 38: 211-222.

Galeano, C. H., Fernandez, A. C., Franco-Herrera, N., Cichy, K. A., McClean, P. E., Vanderleyden, J., & Blair, M. W. 2011. Saturation of an intra-gene pool linkage map: towards a unified consensus linkage map for fine mapping and synteny analysis in common bean. (T. Yin, Ed.) PloS one, 6(12).

Galeano, C. H., Fernández, A. C., Gómez, M., & Blair, M. W. 2009. Single strand conformation polymorphism based SNP and Indel markers for genetic mapping and synteny analysis of common bean (Phaseolus vulgaris L.). BMC genomics, 10, 629.

Gill, N., Findley, S., Walling, J. G., Hans, C., Ma, J., Doyle, J., Stacey, G., et al. 2009. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant physiology, 151, 1167–74.

Gutenkunst, R., R.D. Hernandez, S.H. Williamson and C.D. Bustamante. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data, PLOS Genetics 5:e1000695.

Haas, B. J., Delcher, A. L., Wortman, J. R., & Salzberg, S. L. 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics (Oxford, England), 20: 3643–6.

Ibarra-Perez F.J., B. Ehdaie and J.G. Waines. 1997. Estimation of outcrossing rate in common bean. Crop Sci 37: 60–65.

Iwata, A., A. Tek; M. Richard, B. Abernathy, A. Fonsêca, J. Schmutz, N. Chen, V. Thareau, G. Magdelenat, Y. Li, M. Murata, A. Pedrosa-Harand, V. Geffroy, K. Nagaki, S.A. Jackson. 2013. Identification and characterization of functional centromeres of common bean. Plant J. DOI: 10.1111/tpj.12269.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A, Mardis, E. R., Ding, L., & Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22: 568-576.

Kumar A, Bennetzen JL. 1999. Plant retrotransposons. Annu Rev Genet. 33:479-532.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature 436:793–800.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23: 1026-8. PMID: 17309896

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., et al. 2009. Circos: an information aesthetic for comparative genomics. Genome research, 19, 1639–45.

Lavin, M., Herendeen, P. & Wojciechowski, M. Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Syst. Biol **54**, 575–594.

Lippman Z, Martienssen R. 2004. The role of RNA interference in heterochromatic silencing. Nature 431:364-370.

Nature Genetics: doi:10.1038/ng.3008

Lupas, A., M. Van Dyke and J. Stock. 1991. Predicting coiled coils from protein sequences. Science 252: 1162-1164.

Li, H., & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E.S. & Zhang, Z. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397-2399.

Mafi Moghaddam, S., Song, Q., Mamidi, S., Schmutz, J., Lee, R., Cregan, P., Osorno, J. & McClean, P. E. 2013. Developing market class specific InDel markers from next generation sequence data in *Phaseolus vulgaris* L. Frontiers in Plant Science 1: 0.

McClean, P. E., Mamidi, S., McConnell, M., Chikara, S. & Lee, R. 2010. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. BMC Genomics. 11:184.

McClean, P. E., Terpstra, J., McConnell, M., White, C., Lee, R., & Mamidi, S. 2012. Population structure and genetic differentiation among the USDA common bean (Phaseolus vulgaris L.) core collection. Genetic Resources and Crop Evolution, 59: 499-515.

Nei, M., & Li, W. H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences, 76: 5269-5273.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature. 457:551-556.

Pérez-Rodríguez, P., Riaño-Pachón, D. M., Corrêa, L. G. G., Rensing, S. A., Kersten, B., & Mueller-Roeber, B. 2010. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic acids research *38*, D822-D827.

Pritchard, J. K., Stephens, M., & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics, 155: 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M. & Sham, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81: 559-575.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream and B. Barrell. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16: 944-945.

Rubin, C. J., Megens, H. J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, O., Jern, P., Jorgensen, C. B., Archibald, A. L., Fredhold, M. Groenen, M. A. M. & Andersson, L. 2012. Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences, 109: 19529-19536.

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka. S., Hallbook, F., Besnier, F. Carlborg, O., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K. & Andersson, L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature, 464: 587-591.

Nature Genetics: doi:10.1038/ng.3008

SAS Institute. (Cary, NC., 2002).
Scheet, P., & Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics, 78: 629-644.
Schlueter, J. A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J. J., & Shoemaker, R. C. (2004, October 15). Mining EST databases to resolve evolutionary events in major crop species. Genome / National Research Council Canada = Génome / Conseil national de recherches Canada. NRC Research Press Ottawa, Canada. doi:10.1139/g04-047Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, et al. .2011.. Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat Biotechnol. 30:83-89.
Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature. 463:178-183.
Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123: 585-595.
TIHC (The International HapMap Consortium). 2005. A haplotype map of the human genome. Nature 437: 1299-1320.
Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 2010 26: 873-881.
Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256-276.
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973-982.
Witte CP, Le QH, Bureau T, Kumar A. 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98: 13778–13783.
Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D., Li, R., Cheng, T., Jiang, T., Becquet, C., Xu, X., Liu, C, Zha, X., Fan, W., Lin, Y., Shen, Y., Jiang, ., Jensen, J., Hellmann, I, Tang, S., Zhao, P., Xu, H., Yu, C., Zhang, G., Li, J., Cao, J., Liu, Sh. He. N., Zhou, Y., Liu, H., Zhao, J., Ye. C. Du, Z., Pan, G., Zhao, A., Shao, H., Zeng, W., Wu, P., Li, C., Pan, M., Li, J., Yin, X., Li, D., Wang, J., Zheng, H., Wang, W., Zhang, X., Li, S., Yang, H., Lu, C., Nielsen, R., Zhou, Z., Wang, J. Xiang, Z. & Wang, J. 2009. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). Science, 326: 433-436.
Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. & Buckler, E. S. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38: 203-208.

Nature Genetics: doi:10.1038/ng.3008